NTNU
Norges teknisk-naturvitenskapelige universitet
Institutt for marin teknikk

MASTEROPPGAVE

Tittel:	**Innlevert:**
Sammenligning av ulike metoder innen ekstremverdistatistikk	10. juni 2010
	Tilgjengelighet: Åpen
Student: Christoffer Bartz-Johannessen	**Antall sider:** 57 (totalt med appendiks, men uten oppgavetekst)

Sammendrag:

(sammendrag finnes også i oppgaveteksten, dette er et utdrag av det)

Hensikten med denne oppgaven har vært å se på ulike metoder som i dag benyttes for å predikere ekstremverdier i marin sammenheng. Fem ulike metoder er sammenlignet. Metodene er valgt på grunnlag av hvor mye de benyttes og hvor gode de er ansett for å være.

Det er sett på ekstremverdier for en fiktiv men realistisk responsstørrelse avhengig av bølgeparametrene H_S og T_P. 100, 1000 og 10 000-års verdier for responsstørrelsen er beregnet ved de ulike metodene. Utgangspunktet for alle beregningene er en ca 51 år lang tidsserie med verdier av H_S og T_P.

De resulterende ekstremverdiene beregnet ved de ulike metodene er forholdsvis like. Men noen av metodene er allikevel å foretrekke fremfor andre. Hvor god de ulike metodene er, er vurdert ut fra; hvor gode antagelsene metodene bygger på er, hvor godt metodene benytter informasjonen i de tilgjengelige dataene, hvordan metodene tar hensyn til stor avhengighet i dataene, og hvor enkel metodene er å bruke.

Den beste metoden av de det her er sett på, er etter forfatters mening, Næss-Gaidai-metoden hvor prinsippet går ut på å modellere nivåoppkrysningsfrekvensen til prosessen man studerer. Det er ved denne metoden antatt at antall ganger prosessen krysser et høyt nivå iløpet av en tidsperiode T, er Poissonfordelt, noe som er en god antagelse, med parameter lik nivåoppkrysningsfrekvensen multiplisert med T. Ved å modellere nivåoppkrysningsfrekvensen kan man derfor finne det nivået som krysses med ønskelig returperiode.

Tre Stikkord:	**Veileder:**
Ekstremverdistatistikk	Dag Myrhaug og Bernt Leira

Postadresse:
NTNU
Inst. for marin teknikk
N-7491 Trondheim

Besøksadresse:
Marinteknisk Senter
O. Nielsens vei 10

Telefon +47 73 595501
Telefaks +47 73 595697

NTNU Trondheim
Norges teknisk-naturvitenskapelige universitet
Institutt for marin teknikk

HOVEDOPPGAVE I MARIN TEKNIKK

VÅR 2010

FOR

STUD. TECHN. CHRISTOFFER BARTZ-JOHANNESSEN

SAMMENLIKNING AV ULIKE METODER INNEN EKSTREMVERDI-STATISTIKK

Ved dimensjonering av konstruksjoner og skip til havs kreves pålitelige beregninger og vurderinger av belastninger på og bevegelser av konstruksjonen under påvirkning av vind, bølger og strøm. En god beskrivelse av bølgeforholdene krever blant annet en pålitelig statistisk analyse av tilgjengelige værobservasjoner og bølgedata for det området der konstruksjonen skal plasseres og for de havområdene skipet skal operere. Ved dimensjonering ønsker man å prediktere de største krefter og bevegelser (responsstørrelser) som konstruksjonen forventes å bli utsatt for i løpet av levetiden. Det er flere metoder som kan benyttes for å bestemme en dimensjonerende verdi (for eksempel en såkalt 100 års verdi), men felles er at statistiske modeller tilpasses til tilgjengelige måledata. Slike måledata kan være observasjoner av responsstørrelsen selv eller av de størrelsene denne avhenger av. Målet for denne oppgaven er derfor å vurdere ulike metoder som er vanlig å bruke for å prediktere ekstremverdier og å sammenlikne disse metodene.

Studenten skal:

1. Gi en beskrivelse av de ulike metodene som sammenliknes..
2. Gi en beskrivelse av de data som brukes i analysen.
3. Gjennomføre analysen av data ved bruk av de ulike metodene.
4. Sammenlikne og vurdere de ulike metodene.

Kandidaten skal i besvarelsen legge frem sitt personlige bidrag til løsning av de problemer som oppgaven stiller.

Påstander og konklusjoner som legges frem, skal underbygges med matematiske utledninger og logiske resonnementer der de forskjellige trinn tydelig fremgår.

Kandidaten skal utnytte de muligheter som finnes til å skaffe seg relevant litteratur for det problemområdet kandidaten skal bearbeide.

Besvarelsen skal være oversiktlig og gi en klar fremstilling av resultater og vurderinger. Det er viktig at teksten er velskrevet og klart redigert med tabeller og figurer. Besvarelsen skal gjøres så kortfattet som mulig, men skrives i klart språk. Telegramstil skal unngås.

Besvarelsen skal inneholde oppgaveteksten, forord, innholdsfortegnelse, sammendrag, hoveddel, konklusjon med anbefalinger for videre arbeide, symbolliste, referanser og eventuelle vedlegg. Alle figurer, tabeller og ligninger skal nummeres.

Besvarelsen leveres i en original og kopier etter avtale. Dersom kandidaten utfører oppgaven i samarbeid med en bedrift, skal det leveres en ekstra kopi for bedriften uten godtgjørelse for dette. Alle eksempler utstyres med komplette sett av tabeller, tegninger etc. Bakgrunnsmateriale kan innleveres i egen mappe som bilag til originalen.

Faglærer kan kreve at kandidaten etter utlevering av oppgaven skal legges frem for godkjenning en skriftlig plan for gjennomføring av besvarelsen. Planen skal inneholde antatt omfang av EDB-bruk som belastes instituttet. Faglærer skal varsles dersom EDB-bruk overskrider budsjett.

I besvarelsen skal det tydelig fremgå hva som er kandidatens eget arbeid, og hva han har tatt fra andre kilder. Referansene settes opp i egen liste.

Besvarelsen leveres i 2 eksemplar

- i underskrevet stand
- med oppgaveteksten
- heftet
- med tilhørende tegninger og/eller datamaskinutskrifter i egen mappe.

Veiledere: Professor Bernt Leira
Professor Dag Myrhaug

Dag Myrhaug
Hovedveileder

Innlevering: 14.06.2010

Forord

Alle offshore konstruksjoner vil bli utsatt for krefter fra bølger, vind og strøm. Når man skal dimensjonere disse konstruksjonene ønsker man derfor å predikere de største krefter eller bevegelser (responsstørrelser) man kan forvente å få iløpet av levetiden, såkalte ekstremverdier.

I praksis ønsker man typisk å predikere den største verdien til en responsstørrelse man kan forvente å få iløpet av en gitt periode, for eksempel 100 år, og så bruke denne som dimensjonerende verdi. Det er flere måter å gjøre dette på, men felles for alle er at man tilpasser statistiske modeller til måledata man har tilgjengelig. Måledata kan enten være observasjoner av responsstørrelsen selv, eller av variablene denne avhenger av. Det er imidlertid mange utfordringer som dukker opp når man skal predikere ekstremverdier, disse er først og fremst;

- tilgjengelige periode med måledata er mye kortere enn periodene man vil predikere ekstremverdier for
- man operer med svært små sannsynligheter, noe som stiller store krav til numeriske løsningsmetoder
- stor avhengighet mellom etterfølgende måledata

Fordi ekstreme hendelser kan få fatale konsekvenser er det viktig å kunne predikere disse best mulig slik at man kan dimensjonere for disse. **Målet med denne oppgaven har derfor vært å se på ulike metoder som idag benyttes for å predikere ekstremverdier, sammenligne disse, og se hvordan disse ulike metodene håndterer problemene nevnt ovenfor**.

Det er som sagt flere ulike metoder som er utviklet for å predikere ekstremverdier. Fordi de ulike metodene bygger på ulike prinsipper, vil også resultatene trolig bli ulike. Store forskjeller i resultatene vil imidlertid bety at minst en av metodene gir store feilprediksjoner. Økt forståelse for de ulike metodene, og en sammenligning av resultater vil derfor være til stor hjelp når man skal velge hvilke metode å bruke for ekstremverdiprediksjon. Dette har vært hensikten bak denne oppgaven.

I denne oppgaven er det sett på ekstremverdistatistikk for en fiktiv men realistisk responsstørrelse avhengig av bølgeparametrene signifikant bølgehøyde, H_S, og topperiode, T_P. Måledataene som har vært benyttet er en tidsserie med verdier av H_S og T_P. Metodene som det er sett på i denne oppgaven er generelle, og kan benyttes uansett når man ønsker å predikere ekstremverdier ved hjelp av en tidsserie med observerte måledata, og ikke bare i marin sammenheng.

Jeg vil takke veilederne mine, Dag Myrhaug og Bernt J. Leira for god hjelp gjennom hele prosjektet. Jeg vil og takke Sverre K. Haver og Arvid Næss for mye og god hjelp.

……..

Christoffer Bartz-Johannessen, 10. juni 2010

Innholdsliste

Appendiks

Vedlegg på CD:

Matlabfilene, et Excel ark, og datafilen med måledataene som er benyttet i denne oppgaven, samt selve masteroppgaven er vedlagt på CD. Filene på CD'en er:

Master.docx	Wordfil med masteroppgaven
maladata.txt	Tekstfil med tidsserien med verdier av H_S og T_P
scatter.xlsx	Excelfil med scatterdiagram og tilpasset fordeling for H_S og T_P
input_og_scatter.m	Matlabfil som leser dataene fra maledata.txt og lager scatterdiagram
klassisk_langtidsintegral.m	Matlabfil som beregner langtidsfordelingen til X
arlig_maks_en_og_en.m	Matlabfil som simulerer en tidsserie av X og tilpasser Gumbelfordeling til årlige maksima
arlig_maks_gjennomsnitt.m	Matlabfil som simulerer mange tidsserier av X og tilpasser Gumbelfordeling til alle årlige maksima
Ness_en_og_en.m	Matlabfil som simulerer en tidsserie av X og beregner nivåoppkrysningsfrekvensen til hver simulasjon
Ness_gjennomsnitt.m	Matlabfil som simulerer mange tidsserier av X og beregner nivåoppkrysningsfrekvensen til alle simulasjonene
acer.m	Matlabfil som beregner forskjellige oppkrysningsfrekvenser, og gjør et ACER-plot
Tromans_ford_V_og_Z.m	Matlabfil som finner fordeling til V og Z ved Tromans og Vanderschurens metode
Tromans_ekstr_Weibull.m	Matlabfil som beregner ekstremverdier ved Tromans og Vanderschurens metode når Z er modellert med en Weibullfordeling
Tromans_ekstr_Pareto.m	Matlabfil som beregner ekstremverdier ved Tromans og Vanderschurens metode når Z er modellert med en Paretofordeling
FORM_linjer.m	Matlabfil som beregner FORM-linjer
iso_linjer.m	Matlabfil som beregner isolinjer

Sammendrag

Målet med denne oppgaven har vært å se på ulike metoder som idag benyttes for prediksjon av ekstremverdier i marin sammenheng. Fem ulike metoder er sammenlignet. Metodene er valgt på grunnlag av hvor mye de benyttes, og hvor gode de er ansett for å være.

I oppgaven er det sett på en fiktiv men realistisk responstørrelse, kalt X, avhengig av bølgeparametrene signifikant bølgehøyde, H_S, og topperiode, T_P. Utgangspunktet for alle beregningene er en ca 51 år lang tidsserie med observerte[1] verdier av H_S og T_P. For gitte verdier av H_S og T_P er fordelingen til X kjent. Predikerte ekstremverdier for X ved de ulike metodene er sammenlignet. De fem ulike metodene det er sett på i denne oppgaven er:

Metode1, Klassisk langtidsanalyse: Man finner fordelingen til H_S og T_P. Fordi fordelingen til X gitt H_S og T_P er kjent kan man finne marginalfordelingen til X. Denne brukes så til å predikere ekstremverdier for X.

Metode 2, Gumbeltilpasning til årlige maksima: Ved hjelp av en simulert tidsserie av X (Monte Carlo) plukker man ut årlige maksima og tilpasser en Gumbelfordeling til disse. Med denne fordelingen kan man predikere ekstremverdier.

Metode 3, Næss-Gaidai-metoden: Man modellerer her nivåoppkrysningsfrekvensen til prosessen, dvs. tidsserien X. Det er antatt at antall ganger prosessen X krysser et høyt nivå iløpet av tiden T, er Poissonfordelt med parameter lik nivåoppkrysningsfrekvensen multiplisert med T. Man kan da, når man har en modell for nivåoppkrysningsfrekvensen, finne det nivået som krysses gjennomsnittlig en gang iløpet av ønsket returperiode.

Metode 4, Tromans og Vanderschurens metode: Man tilpasser en fordeling til mest sannsynlig største verdi av X i hver storm. En storm defineres som en periode der H_S er over en viss grenseverdi. Man tilpasser også en fordeling til forholdet mellom største observerte verdi av X og mest sannsynlige største verdi av X i hver storm. Ved hjelp av disse fordelingene kan man finne marginalfordelingen til største observerte X i en storm, og man kan da finne største stormrespons iløpet av en ønsket returperiode.

Metode 5, Konturlinjer: Ved denne metoden finner man den verste kombinasjonen til variablene X er avhengig av. Dette gjøres ved såkalte konturlinjer, her ved FORM-linjer. Den verste kombinasjonen av H_S og T_P man kan forvente å få iløpet av q år finnes, og benyttes så i fordelingen til X. 0,9 kvantilen, $X_{0,9}$, benyttes som q-års ekstremverdi.

Hver av disse metodene har sine sterke og svake sider, og tar i ulik grad hensyn til vanskeligheter med ekstremverdipredikering som; kort mengde data i forhold til perioden man vil predikere ekstremverdier for, og stor avhengighet i måledataene (særlig for værdata slik som i denne oppgaven).

Ekstremverdier for responsstørrelsen, X, beregnet ved de ulike metodene er grovt sett ganske like. I tabellen nedenfor er disse vist for et tilfelle der X er en responstørrelse kvadratisk avhengig av H_S og med egenperiode på 30 sekunder. X_{100}, X_{1000} og $X_{10\,000}$ de X-verdiene som overskrides gjennomsnittlig

[1] Tidsserien med verdier av H_S og T_P er hindcastdata og ikke virkelige observerte verdier. Se kapittel 1.2.

en gang iløpet av henholdsvis 100, 1000 og 10 000 år. Verdiene i rutene er ikke så viktig, det er først og fremst forskjellen i verdiene mellom de ulike metodene som er av interesse.

	Klassisk langtidsanalyse	Gumbeltilpasning till årlige maksima	Næss-Gaidai-metoden	Tromans og Vanderschurens metode	Konturlinjer (FORM-linjer)
X_{100}	684	644	675	675	735
X_{1000}	908	825	896	933	937
$X_{10\,000}$	1159	1006	1144	1210	1152

Man kan legge merke til at ekstremverdiene oppnådd med Gumbeltilpasning til årlige maksima er lavere enn ved de andre metodene. Konturlinjemetoden virker å prediker noe høy 100-års verdi, mens Tromans og Vanderschurens metode predikerer noe høy 10 000-års verdi sammenlignet med øvrige metoder. Med tanke på unøyaktigheter i modelltilpasninger er resultatene ellers praktisk talt like.

Hvor god de ulike metodene er, er vurdert ut fra; hvor gode antagelsene metodene bygger på er, hvor godt metodene benytter informasjonen i de tilgjengelige dataene, hvordan metodene tar hensyn til stor avhengighet i dataene, og hvor enkel metodene er å bruke.

Det er i denne oppgaven konkludert med at en Gumbelfordeling ikke alltid passer bra for å modellere årlige maksima. Denne metoden er derfor ikke å foretrekke selv om den er veldig enkel å bruke. Ved Tromans og Vanderschurens metode virker resultatene å være ganske følsom for hvilken fordeling man benytter når man finner fordelingen til mest sannsynlig største X-verdi i hver storm. Ved konturlinjemetoden er det vanskelig å si hvor god 0,9 kvantil antagelsen er uten å sammenligne med andre metoder. På grunn av mye enklere utregninger enn ved klassisk langtidsanalyse er derfor Næss-Gaidai-metoden den metoden som fremstår som den beste og enkleste å benytte av disse metodene det her er sett på. Denne metoden bygger på få og gode antagelser, tar høyde for avhengighet i måledataene, og er enkel og bruke.

1. Innledning

Det er idag flere metoder som benyttes for å predikere ekstremverdier for krefter eller bevegelser (responsstørrelser) på konstruksjoner enten til havs eller på land som utsettes for stokastiske laster. Alle disse metodene har sine sterke og svake sider, og noen metoder vil trolig være bedre enn andre. Målet med denne oppgaven har derfor vært å se på noen av de vanligste og best ansette metodene, og sammenligne disse.

En sammenligning av de ulike metodene vil være nyttig fordi, hvis forskjellene i resultatene er store vil det si at minst en av metodene predikerer med store feil. En bedre forståelse for de ulike metodene vil og være nyttig for å kunne si hvilke metoder som er mer pålitelig enn andre.

For å gjøre en sammenligning er det sett på ekstremverdier for en fiktiv men realistisk responsstørrelse, X, avhengig av bølgeparametrene H_S og T_P. Ved hjelp av en ca 51 år lang tidsserie med observerte[2] verdier for H_S og T_P er ekstremverdier for X ved de ulike metodene beregnet.

I denne oppgaven vil det bli sett på tre ulike ekstremverdier; X_{100}, X_{1000} og $X_{10\,000}$. Dette er de verdiene som overskrides gjennomsnittlig en gang iløpet av henholdsvis 100, 1000 og 10 000 år. Eller sagt på en annen måte, dette er verdiene med returperiode på 100, 1000 og 10 000 år. Ekstremverdier med disse returperiodene er valgt fordi det er disse returperiodene som vanligvis benyttes som dimensjonerende verdier i reelle situasjoner.

De ulike metodene vil bli presentert i kapittel 2, 3, 4, 5 og 6. I kapittel 7 vil alle resultatene bli presentert, og diskutert.

1.1. Utfordringer med ekstremverdistatistikk

Målet med ekstremverdistatistikk er å predikere typisk 100, 1000 eller 10 000 års verdier til en variabel ved hjelp av målte observasjoner. Problemet er at man vanligvis bare har data for en mye kortere periode, i dette tilfellet ca 51 år. Man må derfor tilpasse en fordeling til de tilgjengelige dataene, og håpe at halen til denne fordelingen er riktig.

Et annet problem når det gjelder modellering av værobservasjoner er ”klumping”. Med det menes at været kommer i klumper, og avhengigheten mellom etterfølgende observasjoner er høy. Hvis man, som man vanligvis gjør, tilpasser en fordelingsfunksjon til de observerte måledataene antar man at hver observasjon er uavhengig, dette vil derfor ikke alltid være en god tilnærmelse p.g.a. klumpingen, se kapittel 2.4.

I ekstremverdistatistikk ønsker man å predikere verdier som blir overskredet veldig sjeldent, for eksempel en gang iløpet av 10 000 år. Dette gjør at sannsynlighetene man jobber med er veldig små, gjerne i størrelsesorden 10^{-4} til 10^{-7}. Mange av kalkulasjonene som blir utført blir gjort numerisk, for eksempel integralutregninger som ikke kan løses analytisk. For at resultatene skal bli fornuftige og uten alt for stor unøyaktighet stiller det derfor store krav til de numeriske løsningsmetodene.

[2] Tidsserien med verdier av H_S og T_P er hindcastdata og ikke virkelige observerte verdier. Se kapittel 1.2.

Metoder der man slipper beregningskrevende utregninger som må løses numerisk er derfor å foretrekke.

1.2. Måledata

Måledataene som er utgangspunktet for alle beregninger i dette prosjektet er en tidsserie med verdier av signifikant bølgehøyde H_S, og topperiode T_P. Det er viktig å presisere at dette ikke er virkelige observerte verdier av bølgeparametrene, men hindcastdata, altså verdier beregnet ved hjelp av andre metrologiske observasjoner / målinger. Tidsserien med verdier av H_S og T_P som er benyttet i denne oppgaven er laget på følgende måte: Basert på målinger av trykk lages isobarer, som igjen brukes til å beregne vindhastigheter. Disse kan så benyttes videre til å beregne bølgeparametre, som i dette tilfellet H_S og T_P. Verdier for H_S og T_P er beregnet hver tredje time fra september 1957 og ut desember 2008. Området de er beregnet for er i Norskehavet, nærmere bestemt 65,3 grader nord og 7,3 grader øst. Tidsserien med hindcastdataene er gitt av Statoil ASA ved S. Haver, og målingene er utført av Meteorologisk institutt.

Hvert verdipar av H_S og T_P i tidsserien representerer en sjøtilstand, det vil si en 3-timers periode der H_S og T_P er konstant. Signifikant bølgehøyde, H_S, er gjennomsnittet av de en tredjedel høyeste bølgene i en gitt periode, og har enhet meter. Topp periode, T_P, er bølgeperioden for den bølgen med høyest energi / amplitude i en sjøtilstand og er derfor den perioden der bølgespekteret har sin maksimale verdi. T_P har enhet sekund. Både H_S og T_P er langsomtvarierende størrelser iforhold til vanlige responsstørrelser som for eksempel kraft eller bevegelse. Man regner derfor at H_S og T_P er konstant over en periode på tre timer (en sjøtilstand) mens X varierer i denne perioden.

På grunn av at tidsserien med H_S og T_P benyttet i denne oppgaven er hindcastdata og ikke reelle observasjoner, er det og en usikkerhet knyttet til hvor nøyaktig disse dataene er. I denne oppgaven er det imidlertid sett bort ifra denne usikkerheten. I og med at alle metodene benytter seg av den samme tidsserien, og fordi målet her har vært å se på de ulike metodene er det antatt at dette ikke er noe problem. I noen sammenhenger blir og tidsserien med verdier av H_S og T_P omtalt som observerte verdier, selv om de egentlig er hindcastdata.

1.3. Responsstørrelsen, X

Alle ekstremverdianalysene i denne oppgaven er gjort for en responsstørrelse, kalt X, avhengig av H_S og T_P. X er den "største verdien" iløpet av en sjøtilstand, dvs. en 3-timers periode med konstant H_S og T_P. Med "største verdi" menes den største verdien til for eksempel en kraft eller en bevegelse iløpet av tre timer. I en gitt sjøtilstand er det antatt at X følger fordelingen (Haver og Bergstrøm (2009)):

$$F_{X|H_S,T_P}\left(x \mid H_S, T_P\right) = \exp\left(-\exp\left(-\frac{x-\alpha\left(H_S,T_P\right)}{\beta\left(H_S,T_P\right)}\right)\right) \quad (1)$$

Ligning (1) gir fordelingen til en generell responsstørrelse avhengig av bølgeparametrene signifikant bølgehøyde, H_S, og topperiode, T_P. Ved å variere funksjonene α og β kan X representere en rekke ulike reelle responsstørrelser.

Responsstørrelsen, X, som er utgangspunktet for denne oppgaven er fiktiv, dvs. X er ikke en spesifikk responsstørrelse til et spesifikt problem, men funksjonene α og β i ligning (1) og (2) er valgt slik at X godt kan representere et realistisk problem. Grunnen til at det her er benyttet en fiktiv responsstørrelse er kun for å spare tid i og med at hensikten med denne oppgaven har først og fremst har vært å se på de statistiske metodene som benyttes for å predikere ekstremverdier. Selv om X derfor ikke er en virkelig responsstørrelse kan den allikevel representere et realistisk problem, og prediksjon av ekstremverdier er like interessant. Det er og antatt at fordelingen til X gitt i ligning (1) er "den riktige", altså ingen usikkerhet rundt denne.

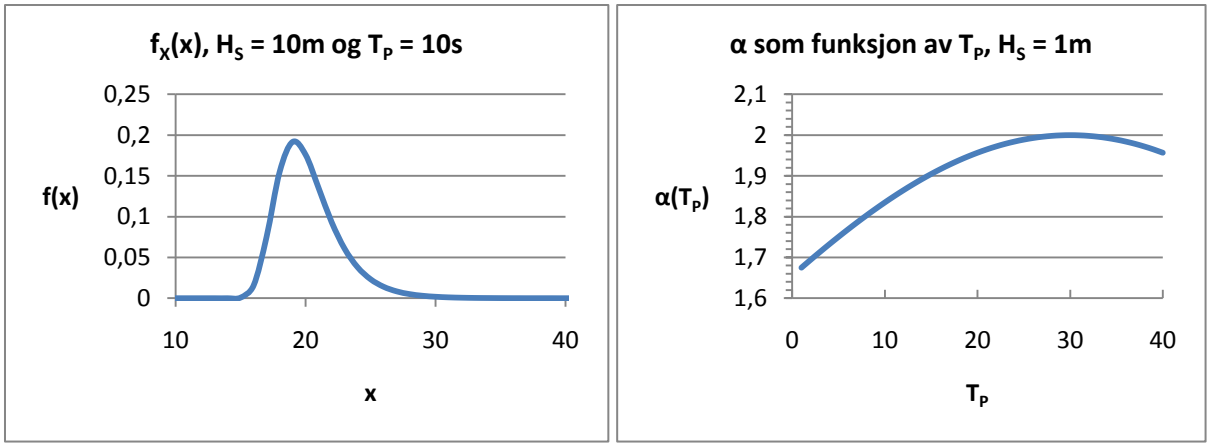
Som man ser av ligning (1) er største respons iløpet av en sjøtilstand, X, Gumbelfordelt med parametre α og β. Lokasjonsparameteren, α, er den mest sannsynlige verdien, mens β er en skala parameter. Begge disse parametrene er avhengig av sjøtilstanden, dvs. H_S og T_P, og er gitt ved (Haver og Bergstrøm (2009)):

$$\alpha\left(H_S,T_P\right)=H_S{}^{\psi}\left[1+\cos^{r}\left(\frac{2\pi\left(T_P-t_0\right)}{p}\right)\right] \qquad (2)$$
$$\beta\left(H_S,T_P\right)=\lambda\cdot\psi\cdot\alpha\left(H_S,T_P\right)$$

Parametrene i ligningene for α og β kan justeres slik at fordelingen til responsstørrelsen X kan tilpasses ulike reelle situasjoner. Ved å sette Ψ = 1 får man en responsstørrelse som varierer lineært med H_S, mens hvis man setter Ψ = 2 får man en responsstørrelse som varierer kvadratisk med H_S. Ved å tilpasse r og p kan man og justere hvor følsom responsen er for eksitasjon nær egenperioden, t_0. Parameteren λ avgjør hvor smal fordelingen til X er, det vil si variabiliteten til X for gitt H_S og T_P.

I denne oppgaven er parameterverdiene i ligning (2) satt til; t_0 = 30s, r = 2, p = 300, og λ = 0,1. Dette gir en respons som er middels følsom for eksitasjon nær egenfrekvensen. Det er sett på tre ulike verdier for Ψ; Ψ = 1, Ψ = 1,5 og Ψ = 2. Dette gir tre ulike responsproblemer, et lineært problem når Ψ = 1, et kvadratisk problem når Ψ = 2, og en mellomting når Ψ = 1,5. Disse verdiene for Ψ er bevist valgt fordi, i reelle situasjoner vil vanligvis responsen være avhengig av bølgehøyden opphøyd i en faktor på mellom 1 og 2.

Verdien til $\alpha(H_s,T_p)$ er plottet som funksjon av T_P til høyre i figur 1 når Ψ = 1 og H_S = 1m for å vise hvordan $\alpha(H_s,T_p)$ varierer for ulike verdier av T_P. Fordelingen til X, for Ψ = 1, H_S = 10m og T_P = 10s, er plottet til venstre i figur 1. Man kan legge merke til at variasjonen til X er forholdsvis stor i en gitt sjøtilstand.



Figur 1: Venstre: fordeling til X for Ψ = 1, H_S = 10m og T_P = 10s. Høyre: α(H_s,T_p) som funksjon av T_P når H_S = 1m, verdien til de resterende parametrene er som nevnt ovenfor.

I de videre kapitlene vil utregningene for de ulike metodene, for enkelhetsskyld, kun bli presentert for Ψ = 1. Utregningene for alle de tre verdiene av Ψ vil imidlertid bli utført, og resultatene vil samlet bli presentert i kapittel 7.

2. Klassisk langtidsanalyse

2.1. Langtidsfordeling til X

En vanlig metode som benyttes for å predikere ekstremverdier er å bruke langtidsfordelingen til X, $F_X(x)$. Fordelingen gitt i ligning (1) er kortidsfordelingen til X, det vil si fordelingen til X i en gitt sjøtilstand. Langtidsfordelingen til X finnes ved å integrere ut H_S og T_P, man får da fordelingen til X uavhengig av H_S og T_P.

$$F_X(x) = \int_{h_s} \int_{t_p} F_{X|H_S T_P}(x \mid h_s, t_p) \cdot f_{H_S, T_P}(h_s, t_p) dt_p dh_s \tag{3}$$

Med denne kan man så finne for eksempel 100-års verdier, X_{100}, ved å løse ligning (4). N_{100} er antall 3-timersperioder iløpet av 100 år.

$$F_X\left(x_{100}\right) = 1 - \frac{1}{N_{100}} \tag{4}$$

Vanligvis er det umulig å løse ligning (3) analytisk, og man må bruke numerisk metoder. Fordi man ofte ønsker å beregne svært små sannsynligheter setter dette høye krav til den numeriske løsningsmetoden. For å løse ligning (3) må man først finne fordelingen til H_S og T_P, dette gjøres ved hjelp av tidsserien med verdier for H_S og T_P.

En annen måte å finne langtidsfordelingen til X på kunne vært og simulert (Monte Carlo) en "observert" verdi av X for hver sjøtilstand, og deretter ha tilpasset en fordeling til disse simulerte verdiene. Dette ville ha resultert i nesten samme langtidsfordeling som funnet ved integralet i (3), men ikke helt. Ved å simulere mange verdier av X finner man en langtidsfordeling til X ved å benytte alle de observerte verdiene av H_S og T_P. Hvis man derimot finner langtidsfordelingen til X ved å løse ligning (3) benytter man en fordeling til H_S og T_P slik at man kan bevege seg utenfor de observerte verdiene av H_S og T_P. På denne måten får man med mer variabilitet i den endelige langtidsfordelingen til X.

2.2. Fordeling til H_S og T_P

Ved hjelp av tidsserien med verdier av H_S og T_P kan man finne den simultane fordelingen til disse variablene. Dette kan for eksempel gjøres ved å fylle alle observerte par av H_S og T_P inn i et scatterdiagram.

Scatterdiagrammet for tidsserien benyttet i denne oppgaven er vist i tabell 1 nedenfor. Tallene i hver rute av scatterdiagrammet forteller hvor mange 3-timers perioder tidsserien inneholder der de observerte verdiene av H_S og T_P er i intervallet til den aktuelle ruten.

H_S \ T_P	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20	20-21	21-22	Sum
0-1	3	173	917	1866	3591	2053	1914	1643	1048	497	311	127	60	0	35	19	0	8	0	0	14265
1-2	0	18	983	4474	8364	6343	8188	7855	6110	4559	2879	1443	646	0	334	128	0	51	0	22	52397
2-3	0	0	5	428	3687	3523	4072	5050	6019	5233	3904	2500	1280	0	481	198	0	48	0	15	36443
3-4	0	0	0	3	376	1162	2380	3046	3629	3806	3305	2104	1327	0	549	222	0	51	0	10	21970
4-5	0	0	0	0	7	74	534	1257	2291	2611	2142	1517	1010	0	368	187	0	18	0	2	12018
5-6	0	0	0	0	0	1	29	232	824	1705	1614	936	581	0	260	118	0	7	0	2	6309
6-7	0	0	0	0	0	0	1	21	105	634	1214	706	345	0	176	86	0	6	0	1	3295
7-8	0	0	0	0	0	0	0	1	15	71	599	671	256	0	79	79	0	1	0	0	1772
8-9	0	0	0	0	0	0	0	0	1	5	121	442	229	0	34	39	0	2	0	0	873
9-10	0	0	0	0	0	0	0	0	0	0	10	154	170	0	50	22	0	1	0	0	407
10-11	0	0	0	0	0	0	0	0	0	0	0	27	84	0	27	21	0	0	0	0	159
11-12	0	0	0	0	0	0	0	0	0	0	0	5	32	0	13	6	0	0	0	0	56
12-13	0	0	0	0	0	0	0	0	0	0	0	0	5	0	11	3	0	0	0	0	19
13-14	0	0	0	0	0	0	0	0	0	0	0	0	1	0	6	2	0	0	0	0	9
14-15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	4
15-16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16-17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
sum	3	191	1905	6771	16025	13156	17118	19105	20042	19121	16099	10632	6026	0	2425	1133	0	193	0	52	149997

Tabell 1: Scatterdiagram for H_S [m] og T_P [s].

Som man ser er noen av kolonnene tomme, dette skyldes unøyaktigheter når verdiene for T_P ble beregnet ved hindcastmetoden. Det kan derfor se ut som at T_P er en diskret variabel. Dette er imidlertid ikke riktig, og en kontinuerlig fordeling er tilpasset T_P.

Det er valgt å først finne fordelingen til H_S uavhengig av T_P, deretter finne fordelingen til T_P gitt H_S. Den simultane fordelingen er da gitt ved ligning (11).

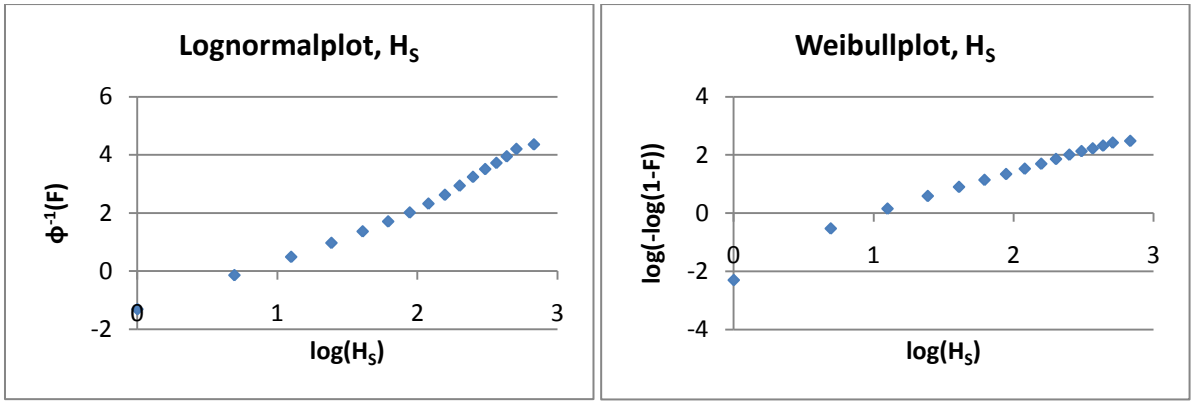
Sannsynlighetspapir er benyttet for å finne fordelingen til H_S. Den empiriske kumulative fordelingen til H_S er beregnet ved hjelp av ligning (5).

$$F_{H_S}(h_S) = \frac{n}{N+1} \qquad n: \text{ antall sjøtilstander } \leq h_S, \quad N: \text{Totalt antall sjøtilstander} \qquad (5)$$

For ulike verdier av H_S leses n av i siste kolonne i scatterdiagrammet (tabell 1). N er totalt antall observasjoner og er siste rute i scatterdiagrammet, dvs. N = 149997.

Ved hjelp av sannsynlighetspapir ser man at nedre del av fordelingen til H_S er veldig nært lognormalfordelt mens øvre del er veldig nært Weibullfordelt, se figur 2. Det er derfor valgt å modellere H_S med en lognormalfordeling for små verdier av H_S, og en Weibullfordeling for store

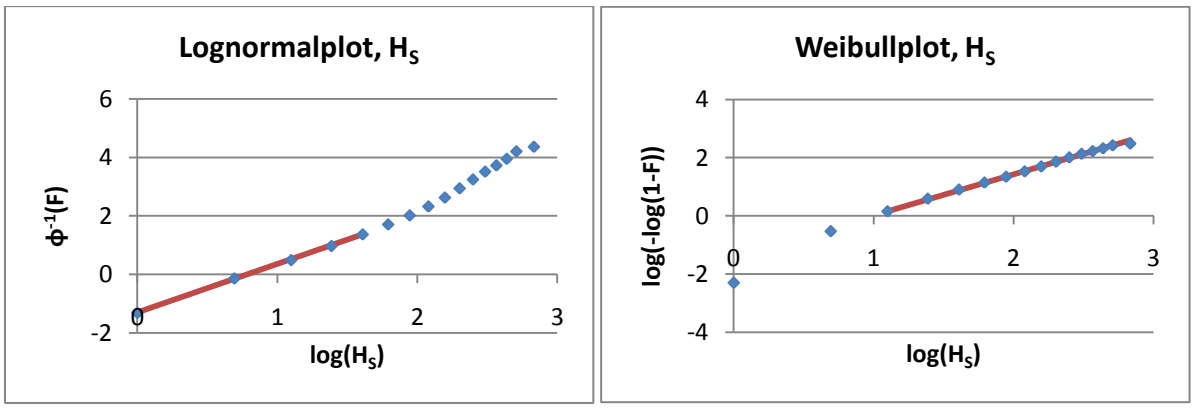
verdier. Dette er og i samsvar med andre modeller for H_S fra nærliggende områder, se for eksempel (Åsgard Metocean Design basis (2004)).



Figur 2: Venstre: lognormalplot for H_S. Høyre: Weibullplot for H_S.

Man ønsker å dele fordelingen til H_S slik at både øvre og nedre fordelinger passer så godt som mulig til observasjonene, samtidig ønsker man at både den kumulative fordelingen, F, og tetthetsfunksjonen, f, skal være kontinuerlig i delingspunktet, dvs. verdien til H_S der man bytter fra lognormal til Weibullfordeling. Å tilfredsstille alle disse tre kriteriene er vanskelig. Det er antatt at det viktigste er at fordelingene er så godt tilpasset dataene som mulig, og at F er kontinuerlig i delingspunktet, noe annet ville vært ufysisk. En diskontinuitet i tetthetsfunksjonene er selvfølgelig ikke ønskelig, men er allikevel mindre alvorlig enn en diskontinuitet i F.

Av figur 2 kan man legge merke til at observasjonene av H_S ligger på en rett linje i et lognormalpapir opp til $H_S \approx 4$ til 5 meter, mens observasjonene ligger på en rett linje i et Weibullpapir fra $H_S \approx 3$ meter og oppover. For best mulig modelltilpasning er det derfor tilpasset en lognormalfordeling til nedre del av dataene, og en Weibullfordeling til øvre del, disse er vist som røde linjer i figur 3. Modelltilpasningene er her gjort enkelt ved å tilpasse en rett linje til observasjonene. Det er antatt at dette er greit nok i og med at punktene ligger på en nærmest perfekt lineær linje.

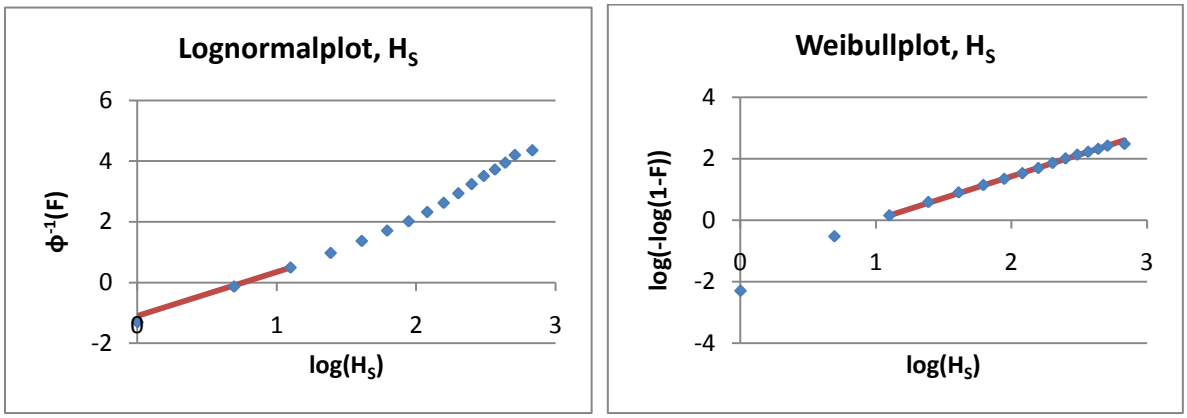


Figur 3: Venstre: lognormalplot for H_S. Høyre: Weibullplot for H_S. Blå prikker er observerte verdier, mens røde linjer er tilpassede fordelinger.

Det er rimelig å prøve og finne et delingspunkt i området der begge fordelingene er gode, dvs. i området $H_S = 3$ til vel 4 meter. Hvis det finnes et punkt i dette området der både F og f er kontinuerlig for de tilpassede fordelingene hadde problemet vært løst. Men dessverre finnes det ikke et slikt

punkt når man benytter fordelingene (de røde linjene i figur 3) funnet over. Med de tilpassede modellene kan man velge om man i delingspunktet vil ha F eller f kontinuerlig, men ikke begge.

Ved å modifisere den tilpassede lognormalfordelingen minimalt, kan man imidlertid få en kontinuerlig overgang for både F og f. Den modifiserte tilpasningen er vist under til venstre i figur 4. Grunnen til at det her er nedre del av fordelingen som er noe modifisert er gjort bevist fordi det er den øvre delen som er viktigst i forbindelse med ekstremverdipredikering.



Figur 4: Endelige modeller for fordelingene til H_S. Delingspunkt er H_S = 3 meter. De tilpassede fordelinger gir kontinuerlig F og f i delingspunktet. Man kan se at den tilpassede modellen for nedre del (i lognormalplottet) er litt dårligere enn i figur 3. Den lille forskjellen har imidlertid ingen praktisk betydning for ekstremverdiprediksjon.

Den lille endringen i modelltilpasningen i nedre del er helt ubetydelig, og har ingen praktisk betydning annet enn at tetthetsfunksjonen ikke vil ha et hakk slik den ville hatt hvis ikke f var kontinuerlig. Ekstremverdiresultatene blir helt like uansett om man benytter lognormalfordelingen i figur 3 eller lognormalfordelingen i figur 4.

Parametrene i de endelige fordelingene til H_S er beregnet enkelt ved å lese av stigningstall og konstantledd til linjene i figur 4. Det er antatt at dette er greit i og med at observasjonene ligger på så godt som perfekte rette linjer. De endelige fordelingene er gitt i ligning (6) og (7) under. Delingspunktet er H_S = 3 meter.

$$f_{H_S}(h_S) = \begin{cases} \dfrac{1}{\sqrt{2\pi}\sigma h_S}\exp\left(-\dfrac{\log(h_S)-\mu^{\,2}}{2\sigma^2}\right) & h_S < 3\text{m} \\ \dfrac{k}{\lambda}\left(\dfrac{h_S}{\lambda}\right)^{k-1} e^{-\left(\frac{h_S}{\lambda}\right)^k} & h_S \geq 3\text{m} \end{cases} \tag{6}$$

Verdiene på parametrene er funnet til:

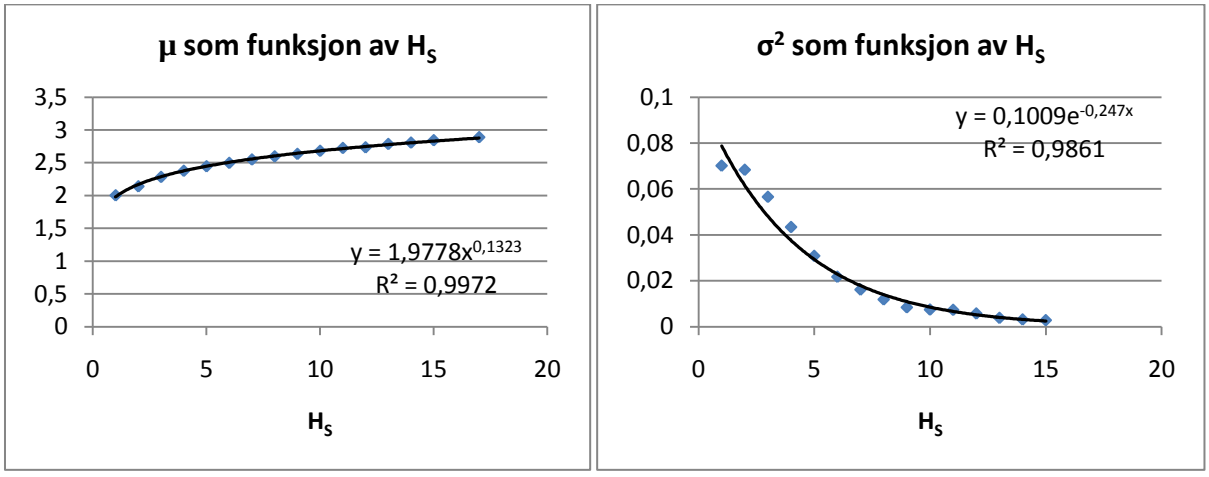
$$\begin{aligned} \mu = 0,761 \;\;\&\;\; \sigma^2 = 0,478 \\ k = 1,412 \;\;\&\;\; \lambda = 2,696 \end{aligned} \tag{7}$$

Figur av fordelingen er vedlagt i appendiks E.

For å finne fordelingen til T_P gitt H_S er det også benyttet sannsynlighetspapir. Hvis man plotter fordelingen til T_P for ulike H_S verdier finner man at fordelingen til T_P for de ulike H_S verdiene vil være veldig nær lognormal. Dette er og i samsvar med andre modeller for T_P gitt H_S benyttet i nærliggende områder, se for eksempel (Åsgard Metocean Design basis (2004)). Parametrene i lognormalfordelingen for hvert H_S intervall finnes ved hjelp av "standard"-formlene for forventning og varians gitt i ligning (8).

$$\mu = E[\log(T_p)] = \frac{1}{n}\sum \log(T_P)$$
$$\sigma^2 = \mathrm{var}(\log(T_p)) = \frac{1}{n-1}\sum (\log(T_P) - \mu)^2 \tag{8}$$

Deretter er μ og σ^2 plottet som funksjon av H_S (se figur 5) for å finne regresjonsmodeller for disse. Disse regresjonsmodellene benyttes så i fordelingen for T_P gitt H_S, man har da fordelingen for T_P gitt H_S for enhver verdi av H_S.



Figur 5: Venstre: μ som funksjon av H_S. Høyre: σ^2 som funksjon av H_S. Sorte linjer er regresjonsmodellene tilpasset i Excel, blå prikker er observerte verdier. I de innfelte formlene for regresjonslinjene er y og x er henholdsvis μ eller σ^2, og H_S.

Med regresjonsmodellene gitt i figur 5 over får man da at fordelingen til T_P gitt H_S blir:

$$f_{T_P|H_S}(t_p \mid h_s) = \frac{1}{\sqrt{2\pi}\sigma(h_s)t_p}\exp\left(-\frac{\log(t_p)-\mu(h_s)^2}{2\sigma^2(h_s)}\right) \tag{9}$$

Med:

$$\mu(h_s) = 1{,}9778h_s^{0{,}1323} \qquad R^2 = 0{,}9972$$
$$\sigma^2(h_s) = 0{,}1009e^{-0{,}247h_S} \qquad R^2 = 0{,}9861 \tag{10}$$

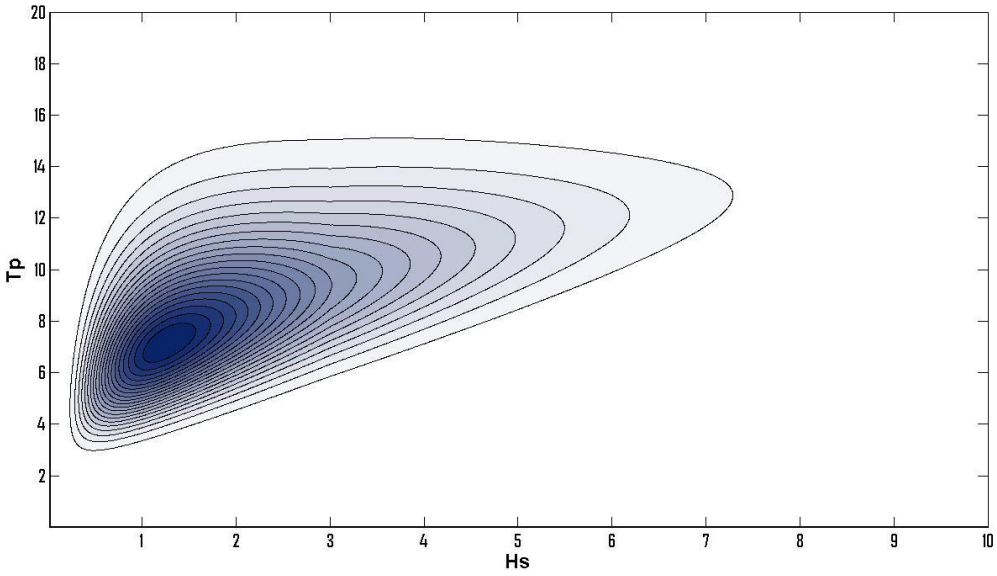
Her er R^2 koeffisientene for regresjonsmodellene, "coefficient of determination", til μ og σ^2. Man må være observant på at disse regresjonsligningene er tilpasset dataene som er samlet inn iløpet av ca 51 år. Når man bruker disse til å predikere for eksempel 10 000-års verdier beveger man seg utenfor

regresjonsmodellenes område, dvs. utenfor der man har data. Man bør derfor være ekstra forsiktig og passe på at regresjonsmodellene oppfører seg "naturlig" også utenfor sitt område.

Den simultane fordelingen til H_S og T_P er da gitt ved:

$$f_{H_S T_P}(h_S, t_P) = f_{H_S}(h_S) \cdot f_{T_P|H_S}(t_P \mid h_S) \tag{11}$$

Den simultane fordelingen er plottet i figur 6 under. Fargen angir verdien til tetthetsfunksjonen.



Figur 6: Tetthetsfunksjonen, $f(h_S,t_P)$, til H_S og T_P med isolinjer. Mørk farge indikerer høy verdi.

2.3. Resultater fra klassisk langtidsanalyse

Med den simultane fordelingen til H_S og T_P kan man løse integralet gitt i ligning (3). Løser man dette integralet numerisk i Matlab kan man finne hvilken X-verdier som gjennomsnittlig overskrides en gang iløpet av 100, 1000 og 10 000 år. Resultatene er gitt nedenfor:

$X_{100} = 38{,}0$

$X_{1000} = 44{,}6$

$X_{10\,000} = 51{,}4$

Dette er ekstremverdiene når $\Psi = 1$. Se tabell 4 og 5 for resultater når $\Psi = 1{,}5$ og $\Psi = 2$.

2.4. Vurdering av metoden

Når man finner den simultane fordelingen til H_S og T_P bruker man hver 3-timers verdi fra tidsserien av H_S og T_P, og tilpasser en fordeling til disse. Men fordi været kommer i klumper er det stor avhengighet mellom nærliggende sjøtilstander, og denne avhengigheten får man ikke med i denne

fordelingen. Dette gjør at resultatene man finner ved å benytte denne fordelingen ikke vil bli helt riktig, men de blir heldigvis konservative. Dette skyldes at ved å bruke verdiene fra hver tredje time vil man for hver storm telle flere "stormer", gitt at stormen varer lenger enn tre timer. Ergo, antallet ekstreme sjøtilstander vil bli talt til et høyere tall enn det virkelige antallet slike hendelser i en tidsperiode. Dette gjør at sannsynligheten for ekstreme sjøtilstander øker. Denne feilen man får p.g.a. avhengigheten mellom nærliggende sjøtilstander ville man også fått om man direkte hadde benyttet en tidsserie av responsstørrelsen X, og tilpasset en langtidsfordeling til denne.

En annen feil man gjør, er å integrere over sjøtilstander som er fysisk umulig. Fordelingen til H_S og T_P er definert for alle positive verdier av H_S og T_P. Det vil imidlertid være fysisk umulig å ha sjøtilstander der H_S er veldig stor og T_P er veldig liten. Volumet til fordelingen i dette "umulige" området er veldig liten, og feilen man innfører her er derfor minimal.

Denne metoden er avhengig av numerisk integrasjon med høy nøyaktighet når man skal beregne ekstremverdiene, dvs. løse integralet gitt i ligning (3). Dette gjør at resultatene kan bli unøyaktige hvis ikke den numeriske løsningsmetoden er nøyaktig nok. Men etterhvert som datakapasiteten generelt blir bedre og bedre vil nok dette problemet bli mindre og mindre.

Å tilpasse en modell ved å bruke alle dataene kan i enkelte tilfeller ikke alltid være et godt alternativ når man vil predikere ekstremverdier. Det kan faktisk være bedre å bruke en mindre mengde av dataene, det er viktigere å bruke all relevant data enn all data (Winterstein (2010)). Hvis man, som i dette tilfellet, har veldig mange observasjoner som man vil tilpasse en fordeling til må man være observant på at halen til fordelingen passer bra til de største av observasjonene. Hvis man benytter for eksempel moment eller sannsynlighetsmaksimeringsestimatorer for å bestemme parametrene er det lett for at fordelingen man finner passer bra der man har mye data, men dårlig der man har lite data, som i halen. Ved å benytte sannsynlighetspapir slik som gjort her er det imidlertid enkelt å sjekke at fordelingen passer bra også i halepartiet. Et annet alternativ kan være og kun benytte de største verdiene i stedet for alle observasjonene.

3. Tilpasning av ekstremverdifordelinger til største verdier

Metoden beskrevet i kapittel 2 er tung utregningsmessig, og man bruker alle de tilgjengelige dataene selv om det først og fremst er de største verdiene som er av interesse. Andre enklere metoder går derfor ut på å kun benytte de største verdiene.

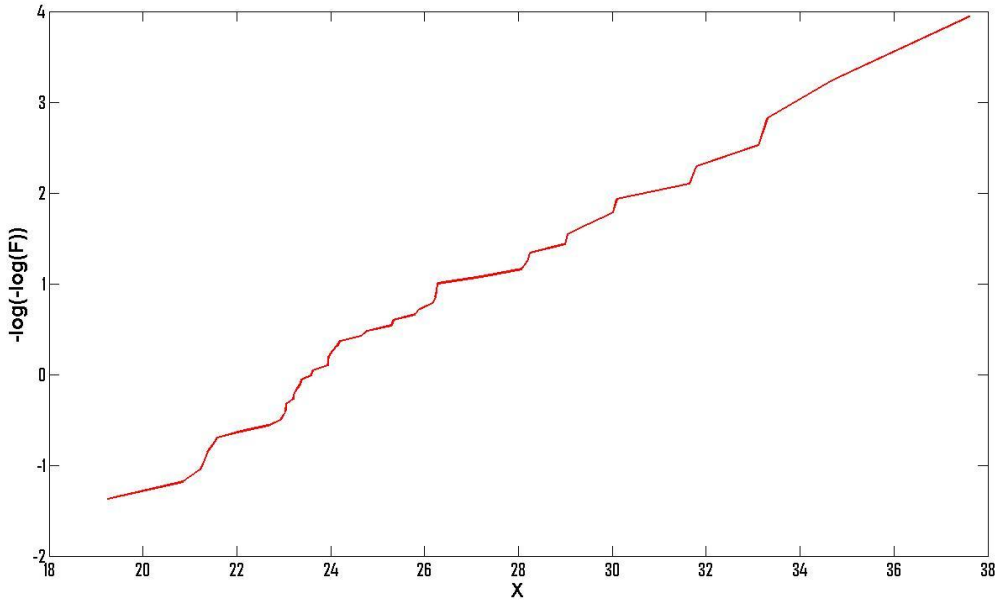
3.1. Gumbelfordeling til årlig maksima

Å tilpasse en Gumbelfordeling til årlige maksima er en enkel metode når man har en tidsserie av variabelen man vil predikere ekstremverdier for. Vi har ikke en observert tidsserie av responsen, X, men i stedet av 3-timers sjøtilstander gitt ved H_S og T_P. Men fordi vi kjenner fordelingen til X i hver av disse sjøtilstandene kan man enkelt lage en "observert" tidsserie av X ved hjelp av Monte Carlo simulering. Fra denne tidsserien kan man så plukke ut årlige maksima til X og tilpasse en Gumbelfordeling til disse. Denne metoden vil i liten grad bli påvirket av avhengigheten mellom nærliggende sjøtilstander, slik som metoden beskrevet i kapittel 2, men den er avhengig i av at hver periode er tilstrekkelig lang slik at andelen av ulike sjøtilstandene er noenlunde riktig representert. Perioder på ett år, slik som benyttet her, burde være lang nok.

Gumbelfordelingen er gitt ved:

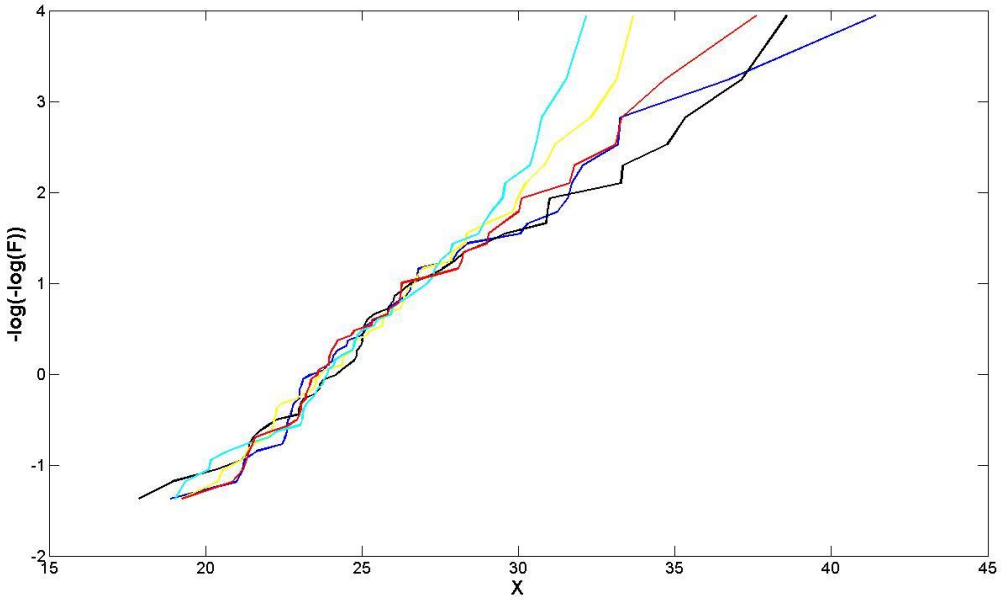
$$F_X(x) = \exp\left(-\exp\left(-\frac{x-\alpha}{\beta}\right)\right) \tag{12}$$

Man kan imidlertid ikke være sikker på om disse årlige maksimumsverdiene vil følge en Gumbelfordeling, se diskusjon appendiks A. Ved å plotte de observerte årlige maksimumene i et Gumbelpapir kan man se hvor god antagelsen er. For å kunne plotte i et sannsynlighetspapir må man beregne verdien til den kumulative fordelingsfunksjonen, F. Dette er gjort som i ligning (5) og det gjelder også i de videre kapitlene. I figur 7 nedenfor er årlige maksima fra en simulasjon av X plottet. Man kan legge merke til at årlige maksima følger forholdsvis bra en rett linje.



Figur 7: Gumbelplot av årlige maksima fra simulert tidsserie av X.

Ved denne metoden benytter man kun årlige maksimalverdier, noe som vil si kun 51 ut av nesten 150000 observasjoner. Fordi man bruker så lite ut av dataene vil resultatene være følsom for variasjon blant de største verdiene. I figur 8 er fem tidsserier av X simulert, og observerte årlige maksimalverdier er plottet i et Gumbelpapir.



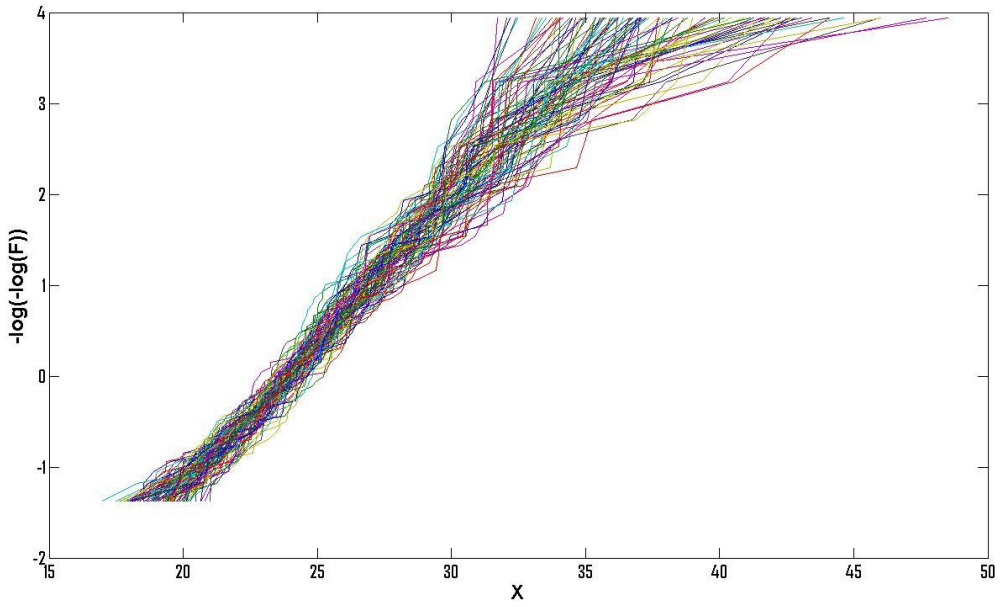
Figur 8: Gumbelplot av årlige maksimum for 5 simulasjoner av X.

Som man ser av figur 8 er variasjonen fra simulasjon til simulasjon forholdsvis stor, i hvert fall med tanke på de største verdiene. Dette skyldes at det er få ekstremt store sjøtilstander i tidsserien av H_S og T_P, derfor vil det være relativt få sjøtilstander som hver gang vil bidra med største verdi iløpet av

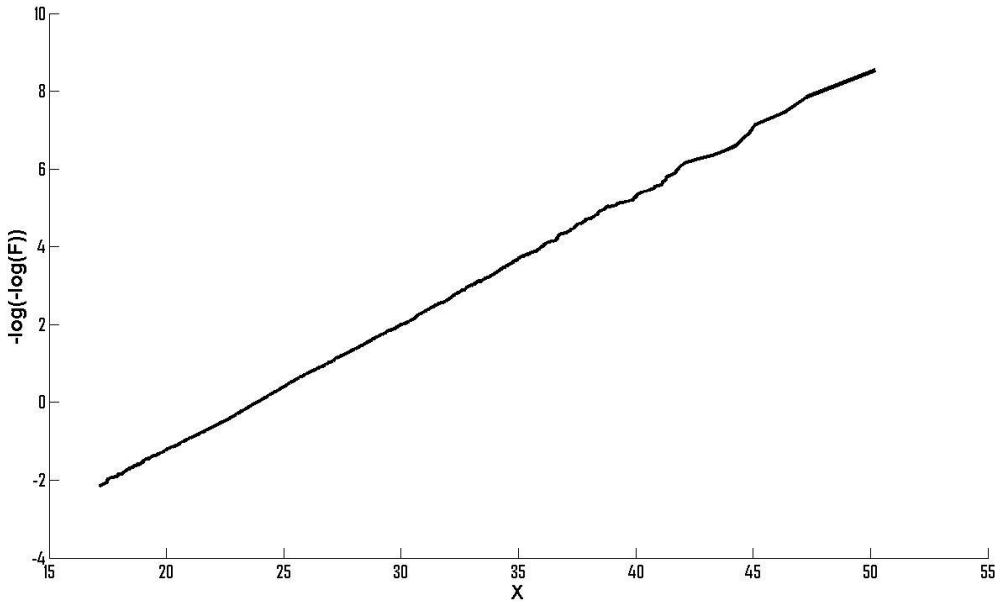
de 51 årene, og fordi fordelingen til X er forholdsvis bred vil variasjonen i største X derfor være ganske stor.

Fordi variasjonen i årlige maksima er ganske stor fra simulasjon til simulasjon, bør man simulere mange tidsserier av X, og benytte alle årlige maksimalverdier. Da vil Gumbelfordelingen man finner i minst mulig grad bli påvirket av tilfeldige variasjoner fra simulasjon til simulasjon. Slik bruker man også mest mulig av informasjonen i tidsserien med verdier av H_S og T_P. På denne måten får man en best mulig fordeling til årlige maksimalverdier gitt tidsserien av H_S og T_P man har tilgjengelig.

For å tilpasse en Gumbelfordeling er 100 tidsserier av X simulert. Alle de 100*51 simulerte verdiene av årlige maksima er benyttet. Disse 51 årlige maksimaene til hver av de 100 simulasjonene er plottet i figur 9. I figur 10 er alle de 100*51 årlige maksimaene plottet sammen.



Figur 9: Gumbelplot av årlige maksimum for 100 simulasjoner av X.



Figur 10: Gumbelplot av 100*51 årlige maksimum.

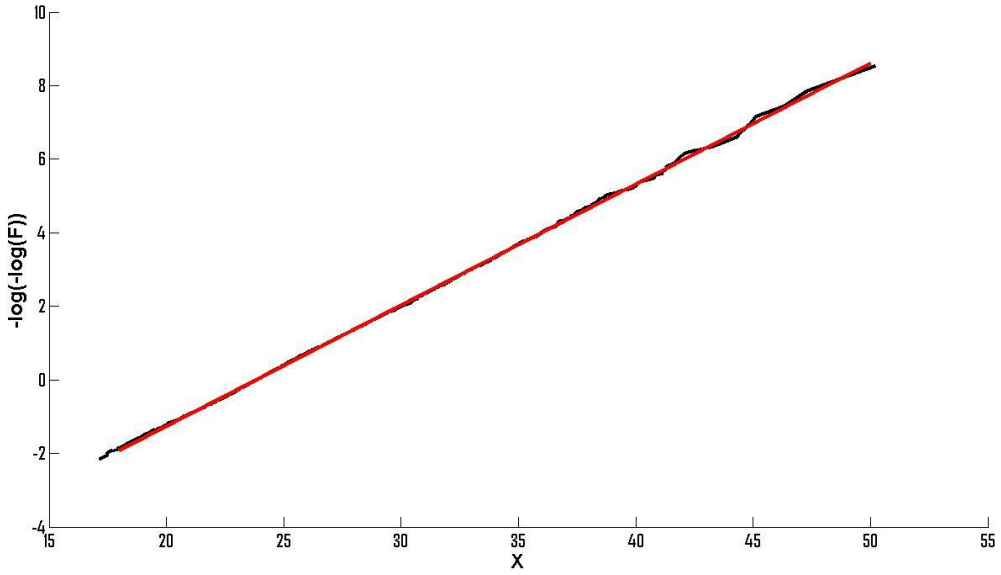
Som man ser av figur 10 følger årlige maksima en ganske fin og lineær linje, man kan derfor anta at en Gumbelfordeling passer godt. Parametrene i fordelingen finnes ved momentmetoden (Haver (2009), Leira (2005)). Dette er gjort i Matlab, resultatene er gitt i ligning (13). Her er μ og σ beregnet forventningsverdi og standardavviket til observasjonene.

$$\begin{aligned} \hat{\beta} &= 0,7797\sigma = 0,7797 \cdot 3,899 = 3,04 \\ \hat{\alpha} &= \mu - 0,57722\beta = 25,59 - 0,57722 \cdot 3,04 = 23,84 \end{aligned} \tag{13}$$

Det kunne vært fristende og heller bare ha tilpasset en linje i Matlab, dette blir da gjort ved minste kvadrats metode. Men for punkt plottet i et Gumbelpapir, eller andre sannsynlighetspapir er ikke dette alltid en god løsning. Dette fordi man her har logaritmiske eller dobbelt logaritmiske skalaer. Når man tar logaritmen til en verdi med en liten feil, vil denne feilen få ulike konsekvenser ettersom hvor på den logaritmiske skalaen man befinner seg. Er for eksempel inputverdien nær null vil en liten feil gi store utslag på verdien til logaritmen, ved store inputverdier vil en liten feil derimot nesten ikke ha noe betydning. Dette gjør at like avvik vil vektes ulikt i en logaritmisk skala, og minste kvadratsmetode er derfor ikke et godt alternativ.

Andre estimatorer kunne og vært benyttet i stedet for momentestimatorene, som for eksempel sannsynlighetsmaksimeringsestimatorer (ML-estimatorer). Utrykkene for ML-estimatorene for Gumbelparametrene er imidlertid ganske kompliserte og må løses numerisk og iterativt. Fordi datasettet som her er benyttet er såpass stort, 100*51 observasjoner, er det trolig ikke så farlig hvilken måte man estimerer parametrene på så lenge estimatorene er konsistente. Det er derfor valgt å benytte momentestimatene for parameterverdiene. Samme begrunnelse gjelder og ved senere parameterestimasjoner for Gumbelparametre.

Figur 11 nedenfor viser den tilpassede fordelingen plottet i et Gumbelpapir (rød linje) sammen med de observerte verdiene (svart linje).



Figur 11: Tilpasset fordelingen (rød linje) sammen med de observerte verdier (svart linje).

3.1.1. Resultater ved Gumbeltilpasning til årlige maksima

Den X-verdien som overskrides gjennomsnittlig en gang iløpet av 100 år finner man ved å løse ligning (14).

$$F_X\left(x_{100}\right)=\exp\left(-\exp\left(-\frac{x_{100}-\hat{\alpha}}{\hat{\beta}}\right)\right)=1-\frac{1}{100} \qquad (14)$$

Dette gir:

X_{100} = 37,8

Tilsvarende finner man den X-verdien som overskrides gjennomsnittlig en gang iløpet av 1000 og 10 000 år.

X_{1000} = 44,8

$X_{10\,000}$ = 51,8

Dette er da resultatene for Ψ = 1. Se tabell 4 og 5 for resultater når Ψ = 1,5 og Ψ = 2.

3.1.2. Vurdering av metoden

En ulempe med denne modellen er den små mengden data man benytter iforhold til hva man har tilgjengelig. I dette tilfellet har ikke dette vært noe problem fordi man har kunnet simulere mange tidsserier av X, og benytte alle årlige maksimalverdier. Hvis man derimot bare hadde hatt en tidsserie av X ville metoden presentert i kapittel 4 vært et mye bedre valg.

Antagelsen om at maksimaene følger en Gumbelfordeling er ikke nødvendigvis riktig, se diskusjon i appendiks A.

Ved å kun plukke ut årlige maksima vil man miste informasjon på den måten at små maksima fra rolige år vil tas med, mens større maksima fra stormfulle år vil ikke bli tatt med. Andre metoder som topp over terskel metoder (engelsk: Peak Over Treshold methods = POT) er bedre fordi alle observasjoner over en viss terskel tas med. Dette er og tanken bak metoden i kapittel 4.

For tilfellene der $\Psi = 1{,}5$ og $\Psi = 2$ passer Gumbelfordelingen mindre bra enn når $\Psi = 1$ som ovenfor. Plotter man årlige maksima for $\Psi = 1{,}5$ er linjen ganske ben, for $\Psi = 2$ derimot er linjen begynt å krumme såpass mye at Gumbelantagelsen ikke er lenger helt god (se figur B i appendiks C). Dette forklarer nok hvorfor resultatene ved denne metoden skiller seg fra resultatene ved de andre metodene for $\Psi = 1{,}5$ og særlig for $\Psi = 2$ i tabell 4 og 5, kapittel 7.

3.1.3. Bakgrunnen for Gumbelantagelsen

Grunnen til at Gumbelfordelingen er valgt for å modellere årlige maksima, og ofte ellers brukes for å modellere ekstremverdier, er at ofte vil den største ut av n uavhengige identisk fordelte variabler følge en Gumbelfordeling når n går mot uendelig. Dette gjelder ikke alle variabler, men det gjelder for eksempel variabler som er eksponential, normal, lognormal, og Rayleigh-fordelt. I og med at de fleste naturlige fenomener, som for eksempel bølgehøyde, kan modelleres ved en av disse fordelingene er det rimelig å anta at den største ut av et stort antall observasjoner vil følge en Gumbelfordeling. Dette er diskutert litt mer detaljert i appendiks A.

Det er imidlertid ikke alltid de årlige maksimumene vil følge en Gumbelfordeling. De ”observerte” X-verdiene er ikke uavhengige (pga at nærliggende sjøtilstander ikke er uavhengig, derfor vil ikke nærliggende X-verdier heller være uavhengig). Om fordelingen til X, $F_X(x)$, er en fordeling der fordelingen til den største verdien, $F_X^n(x)$, vil konvergere mot en Gumbelfordeling er heller ikke sikkert. Ved å plotte årlige maksima i et Gumbelpapir kan man vurdere hvor god Gumbelantagelsen er. Hvis man har begrenset med data kan imidlertid dette være vanskelig.

3.2. Alternative ekstremverdimetoder

En annen ekstremverdimetode som benyttes er topp over terskel metoden. Denne metoden går ut på å benytte alle observasjoner over en viss terskel, u. Det defineres en ny variabel $Y = X – u$. Det kan vises at fordelingen til Y er går mot en generalisert Paretofordeling når u er stor.

Topp over terskel metoden får brukt mer av informasjonen i dataene enn å kun benytte årlige maksima. Men metoden er ikke godt egnet der det er stor avhengighet mellom observasjonene slik at tidsserien klumper seg. Denne metoden er derfor ikke tatt med her fordi metoden i kapittel 4 benytter samme ideen og er samtidig bedre på alle måter.

4. Næss-Gaidai-metoden

Næss-Gaidai-metoden er en metode for å predikere ekstremverdier fra en observert tidsserie som er utviklet av blant andre professor Arvid Næss ved NTNU. Prinsippet bak metoden går ut på å modellere nivåoppkrysningsfrekvensen $\nu^+(x)$ til en prosess X(t) (Næss og Gaidai (2008)). Hvis vi antar at antall ganger prosessen X(t) krysser et nivå x iløpet av tiden T, er Poissonfordelt, da vil parameteren i denne fordelingen være gitt ved $\lambda = \nu^+(x)*T$, der $\nu^+(x)$ er gjennomsnittlig nivåoppkrysningsfrekvens av nivået x. Har man en modell for $\nu^+(x)$ kan man derfor, under Poissonantagelsen, enkelt finne nivået x som gir ønsket antall oppkrysninger iløpet av T.

Man vet at antall hendelser er Poissonfordelt hvis (Løvås (2004)):

- Antall forekomster i disjunkte tidsintervall er uavhengig av hverandre
- Forventet antall forekomster er konstant per tidsenhet
- To forekomster ikke kan inntreffe på samtidig

For høye nivåverdier virker Poissonantagelsen for nivåoppkrysninger å være god (Næss et. al (2009)), altså, antall ganger prosessen krysser et høyt nivå iløpet av en tid T er nært Poissonfordelt. Dette er diskutert i mer detalj i kapittel 4.2.

For en stasjonær prosess, X(t), vet man at nivåoppkrysningsfrekvensen, $\nu^+(x)$, av et nivå x, er gitt ved (Myrhaug (2005)):

$$\nu^+(x) = \int_0^\infty \dot{x} f_{X\dot{X}}(x, \dot{x})\, d\dot{x} = E_+\left[\dot{X} \mid X = x\right] f_X(x) \tag{15}$$

X(t) merket er den tidsderiverte av X(t). Det er antatt at, for store verdier av x, vil $\nu^+(x)$ være på formen (Næss et. al (2009)):

$$\nu^+(x) = q(x)\exp\left(-a(x-b)^c\right) \qquad x > x_1 \tag{16}$$

x_1 er en grenseverdi for hvor ligning (16) gjelder. Ligning (16) gir videre:

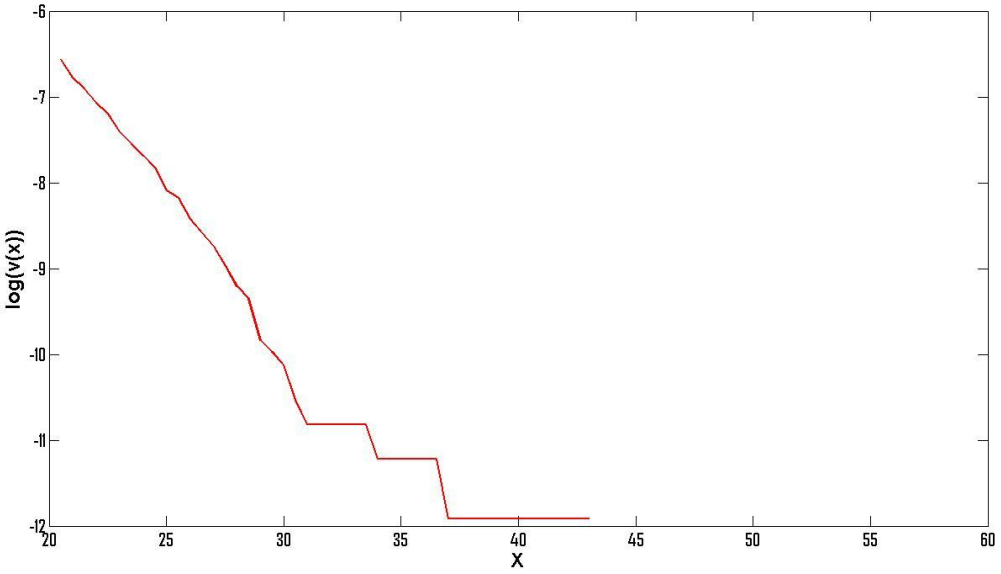
$$\log\left(\nu^+(x)\right) = \log\left(q(x)\right) - a(x-b)^c \qquad x > x_1 \tag{17}$$

Funksjonen q(x) er langsomtvarierende iforhold til $\exp(-a(x-b)^c)$. Man kan derfor anta at q(x) er konstant for store verdier av x (Næss et. al (2009)). Et plot av $\log(\nu^+(x))$ mot x vil indikere verdiene til a, b og c men det er ikke nok til å bestemme de. Flere metoder er foreslått å bestemme verdiene til a, b og c, se for eksempel (Næss et. al (2009)).

Ved hjelp av den simulerte tidsserien av X(t) kan $\nu^+(x)$, for ulike verdier av x, beregnes. Man teller da antall ganger prosessen X(t) krysser nivået x gitt at forrige verdi av X(t) var under x, og deretter deler man på totalt antall observasjoner minus en, N-1, fordi man ikke teller med første observasjon.

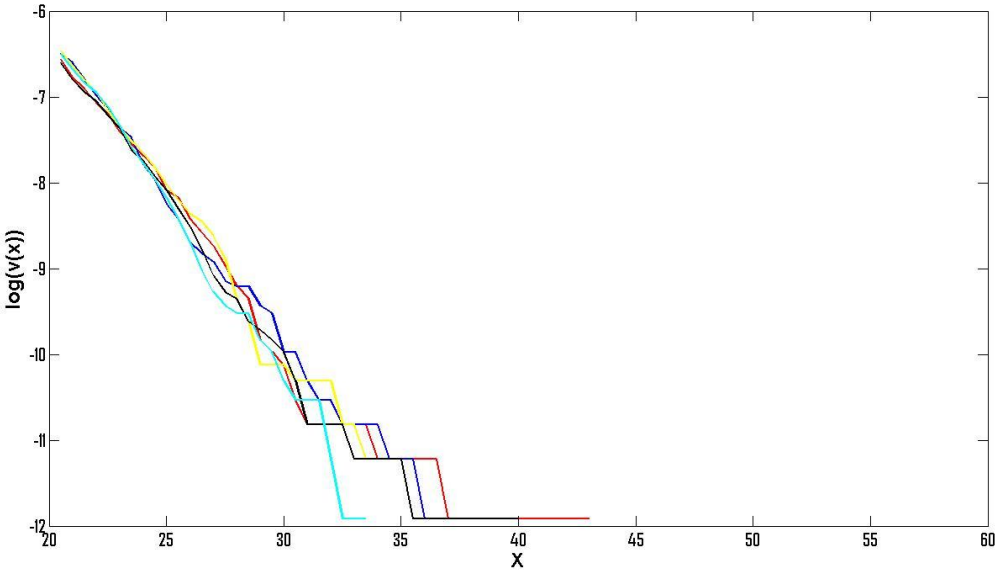
$$\hat{\nu}^+(x) = \frac{1}{N-1}\sum_{i=2}^{N} I_{\{X_{i-1}<x,\, X_i \geq x\}} \tag{18}$$

Ligning (18) viser hvordan den estimerte verdien for $v^+(x)$ beregnes. "I" er indikatorfunksjonen som er 1 bare hvis $X_{i-1} < x$ og $X_i \geq x$. For den simulerte tidsserien av X(t) er $v^+(x)$ beregnet for mange ulike nivåer, x. Logaritmen (den naturlige) til $v^+(x)$ som funksjon av nivået, x, er plottet i figur 12. Her er x_1 er satt til 20.



Figur 12: Logaritmen til $v^+(x)$ til den simulerte tidsserien av X(t) som funksjon av nivået, x.

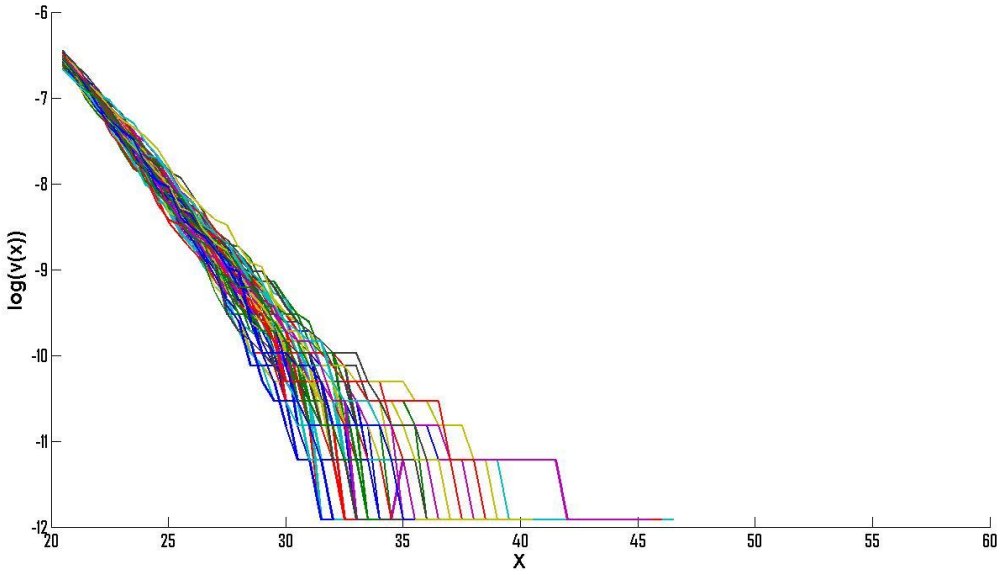
Som for metoden i kapittel 3 vil variasjonen blant de største verdiene være stor. Med denne metoden kan man imidlertid justere x_1 slik at mye mer av de tilgjengelige dataene tas med. Modelltilpasningen blir da mindre utsatt for variasjonen blant de største verdiene. I figur 13 er logaritmen til $v^+(x)$ for fem simulasjoner av X(t) plottet.



Figur 13: Logaritmen til $v^+(x)$ til 5 simulerte tidsserier av X(t) som funksjon av nivået, x.

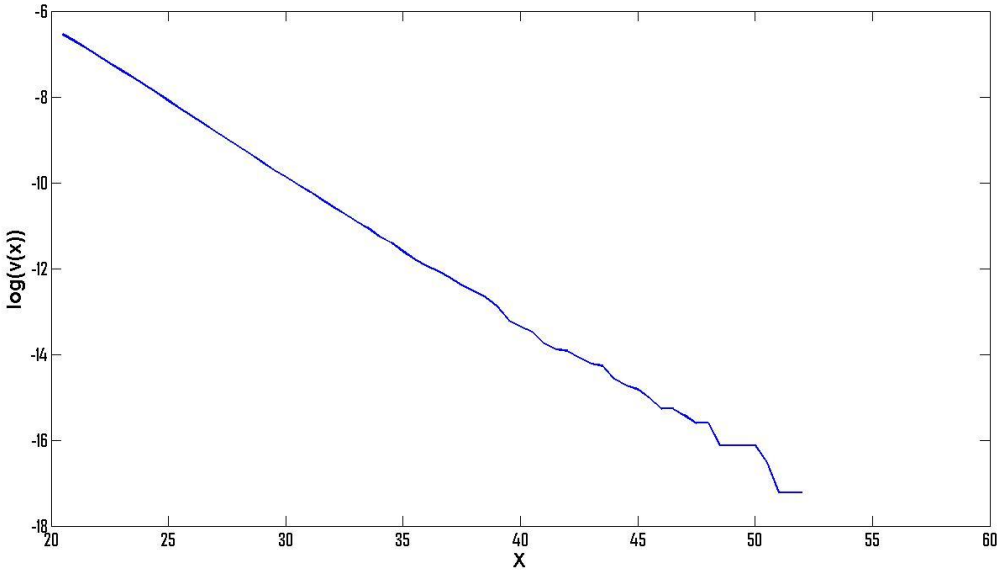
Som i kapittel 3 er det simulert 100 tidsserier av X(t), hele denne ca 100*51 år lange tidsserien av X(t) er brukt for å utnytte mest mulig av informasjonen i den tilgjengelige tidsserien av H_S og T_P.

Logaritmen til $v^+(x)$ for hver av de 100 tidsseriene er plottet i figur 14. Man ser her spredningen i de største verdiene fra simulasjon til simulasjon.



Figur 14: Logaritmen til $v^+(x)$ til 100 simulerte tidsserier av X(t) som funksjon av nivået, x.

Logaritmen til $v^+(x)$ for hele den ca 100*51 år lange tidsserien av X(t) er plottet i figur 15.



Figur 15: Logaritmen til $v^+(x)$ til hele den ca 51*100 år lange tidsserien av X(t) som funksjon av nivået, x.

Som man ser av figur 15 er logaritmen til $v^+(x)$ veldig nær lineær i x. Det er derfor rimelig å anta at c i formel (16) og (17) er tilnærmet lik 1. En funksjon for $\log(v^+(x))$ på formen gitt i (17) finnes i Matlab ved hjelp av kurvetilpasningsverktøy, og er gitt i ligning (19). For $\Psi = 1{,}5$ og $\Psi = 2$ vil imidlertid

log($v^+(x)$) som funksjon av x krumme svakt utover (se figur C appendiks C). For å finne et utrykk for $v^+(x)$ er det også her benyttet kurvetilpasningsverktøy i Matlab for $v^+(x)$ er på formen som gitt i ligning (16).

4.1. Resultater ved Næss-Gaidai-metoden

Formel for log($v^+(x)$) er estimert til:

$$\log\left(\hat{\nu}^+(x)\right) = -0,351x + 0,659 \tag{19}$$

Som gir:

$$\hat{\nu}^+(x) = \exp\left(-0,351x + 0,659\right) \tag{20}$$

Under Poissonantagelsen vil antall ganger prosessen X(t) krysser et nivå x iløpet av tiden T være Poissonfordelt med parameter $\lambda = v^+(x)*T$. Antall krysninger iløpet av 100 år er derfor Poissonfordelt med parameter $\lambda_{100} = v^+(x)*T_{100}$, hvor T_{100} her er det samme som N_{100}, antall observasjoner iløpet av 100 år. For en Poissonprosess er forventet antall hendelser lik parameteren, $\lambda = v^+(x)*T$. Man kan da enkelt finne det nivået, x, som krysses gjennomsnittlig én gang per 100 år ved å løse ligning (21).

$$\lambda_{100} = \nu^+(x) \cdot T_{100} = 1 \tag{21}$$

Fyller vi inn ligning (20) for $v^+(x)$ i ligning (21) og løser for x, får man $X_{100} = 37,73$.

Tilsvarende kan vi finne 1000 og 10 000-års verdier. Disse blir:

$X_{1000} = 44,29$

$X_{10\,000} = 50,85$

Dette er resultatene for $\Psi = 1$. Se tabell 4 og 5 for resultater når $\Psi = 1,5$ og $\Psi = 2$.

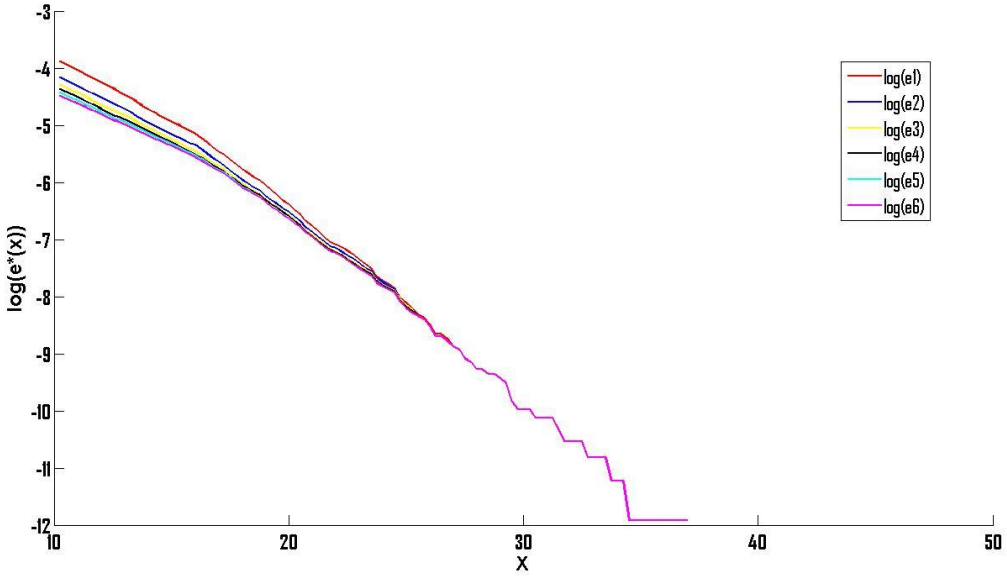
4.2. Poissonantagelsen

For å sjekke hvor god Poissonantagelsen er kan man gjøre et ACER-plot (ACER = average conditional exceedance rate), (Næss og Gaidai (2008)). Når vi plottet nivåoppkrysningsfrekvensen som funksjon av nivået, x, ble denne beregnet ved hjelp av ligning (18). Tilsvarende kan man beregne frekvensen prosessen krysser nivået x, gitt at de to foregående observasjonene var under x. Denne frekvensen kalles $e_2(x)$ og beregnes som i ligning (22).

$$e_2(x) = \frac{1}{N-2}\sum_{i=3}^{N} I_{\{X_i \geq x, X_{i-2} < x, X_{i-1} < x\}} \tag{22}$$

Man kan og finne $e_3(x)$ som er frekvensen prosessen krysser nivået x med, gitt at de tre foregående observasjonene var under x. Tilsvarende kan man finne $e_4(x)$, $e_5(x)$ osv. Plotter man disse forskjellige

oppkrysningsfrekvensene vil man se at de konvergerer for store verdier av x. I figur 16 nedenfor er e_1, $e_2(x)$, opp til $e_6(x)$ for en simulert tidsserie av X(t) beregnet og plottet.



Figur 16: ACER-plot. e*(x) er her en fellesbetegnelse på $e_1(x)$, $e_2(x)$, osv. Man ser at de ulike frekvensene konvergerer for store verdier av x.

I figur 16 ovenfor tilsvarer $e_1(x)$ nivåoppkrysningsfrekvensen, $v^+(x)$, beregnet ved ligning (18). Man ser at for store verdier av x konvergerer de seks frekvensene som her er plottet. Dette betyr at for høye nivå er nivåkrysningene uavhengige, det er ingen "klumping". Hadde det vært klumping ville for eksempel $e_6(x)$ vært mye lavere enn $e_1(x)$, men det er den ikke. Man kan derfor konkludere med at nivåkrysningene på høye nivå er uavhengige. Det er og fornuftig å anta at forventet antall hendelser (nivåkrysninger) iløpet av et rimelig tidsintervall, T, er konstant. Poissonantagelsen er derfor antatt å være god.

4.3. Vurdering av metoden

Næss-Gaidai-metoden bruker mye mer av informasjonen i dataene, gitt at x_1 ikke er for høy, enn ved Gumbeltilpasning til årlige maksima. I stedet for å bare benytte periodiske maksima, benyttes alle hendelser der prosessen krysser et vist nivå. Metoden har heller ikke problemer med "klumping" som i POT metoden fordi man her kun ser på oppkrysningsfrekvensen, altså kun de gangene prosessen $X_i(t)$ er over x gitt at $X_{i-1}(t)$ er under x. Man teller derfor antall klumper over ett nivå, x. Metoden er også enkel å bruke, og krever ingen løsninger av tunge numeriske integral som for eksempel metodene i kapittel 2 og 5.

5. Tromans og Vanderschurens metode

Tromans og Vanderschuren har foreslått en annerledes metode for å predikere ekstremverdier (Haver og Bergstrøm (2009)). Ideen bak denne metoden er å tilpasse en modell til stormene man har observert. Dette er her gjort ved å tilpasse en fordeling til mest sannsynlig største respons i hver storm, denne kalles Z, og er det samme som $\alpha(H_S,T_P)$ gitt i ligning (2). Største observerte respons i hver storm kalles Y, dvs. Y er den største verdien til X iløpet av en storm og Z er største α-verdi. Med "storm" menes perioder der H_S er over en viss grense, $H_{S,storm}$. Det er antatt at forholdet V, mellom Y og Z (V = Y/Z) er Gumbelfordelt, og uavhengig av stormintensiteten, dvs. stormens lengde og styrke.

$$F_V(v) = \exp\left(-\left(\exp\left(-\frac{v-\alpha_V}{\beta_V}\right)\right)\right) \tag{23}$$

Ved å først finne fordelingen til Z ved hjelp av tidsserien av H_S og T_P, kan man deretter finne fordelingen til Y ved å løse integralet i ligning (24). Man finner da fordelingen til Y som er største X-verdi iløpet av en storm, og en storm er en periode der $H_S \geq H_{S,storm}$.

$$F_Y(y) = \int_z F_{Y|Z}(y \mid z)\, f_Z(z)\, dz \tag{24}$$

$F_{Y|Z}(y|z)$ finner man ved å bytte ut V med Y/Z i ligning (23):

$$F_{Y|Z}(y \mid z) = \exp\left(-\left(\exp\left(-\frac{y/z-\alpha_V}{\beta_V}\right)\right)\right) \tag{25}$$

Når man kjenner fordelingen til Y kan man estimere de største stormresponsene man kan forvente å få iløpet av en gitt returperiode. For å finne Y-verdien som overskrides gjennomsnittlig en gang per 100 år løser man ligning (26).

$$F_Y(y) = 1 - \frac{1}{N_{storm,100}} \tag{26}$$

Her er $N_{storm,100}$ forventet antall stormer iløpet av 100 år som er lik antall observerte stormer i tidsserien med H_S og T_P delt på lengden av tidsserien (i år) ganget med 100. Tilsvarende finner man forventet antall stormer for 1000 og 10 000 år.

Y-verdien som finnes ved å løse ligning (26) er da stormresponsen som overskrides gjennomsnittlig en gang iløpet av 100 år. Denne Y-verdien brukes som 100-års verdi for X, X_{100}, altså X-verdien som overskrides gjennomsnittlig en gang per 100 år. Ved å gjøre dette "glemmer" man å ta med sannsynligheten for at største respons skal skje utenfor en storm. Denne sannsynligheten er trolig tilnærmet null, feilen man gjør her er derfor veldig liten.

Man kan spørre seg hvorfor man finner fordelingen til største stormrespons på denne noe kompliserte måten. Man kunne jo bare ha plukket ut største maksima i fra hver storm og tilpasset en fordeling til disse. Hadde man simulert mange tidsserier av X kunne man funnet en fin fordeling til Y

på bakgrunn av de stormene tidsserien av H_S og T_P inneholder. Hvis man så hadde gjort dette, da ville man brukt de samme stormene om og om igjen. Hensikten med å finne en parametrisk fordeling til stormene, ved Z, er derfor, som ellers når man tilpasser en fordeling, at man kan bevege seg utenfor de observerte verdiene. Slik kan man ta høyde for "uobserverte stormer", det vil si sjeldne og kraftige stormer som ikke er kommet med i den "korte" tidsserien med verdier for H_S og T_P.

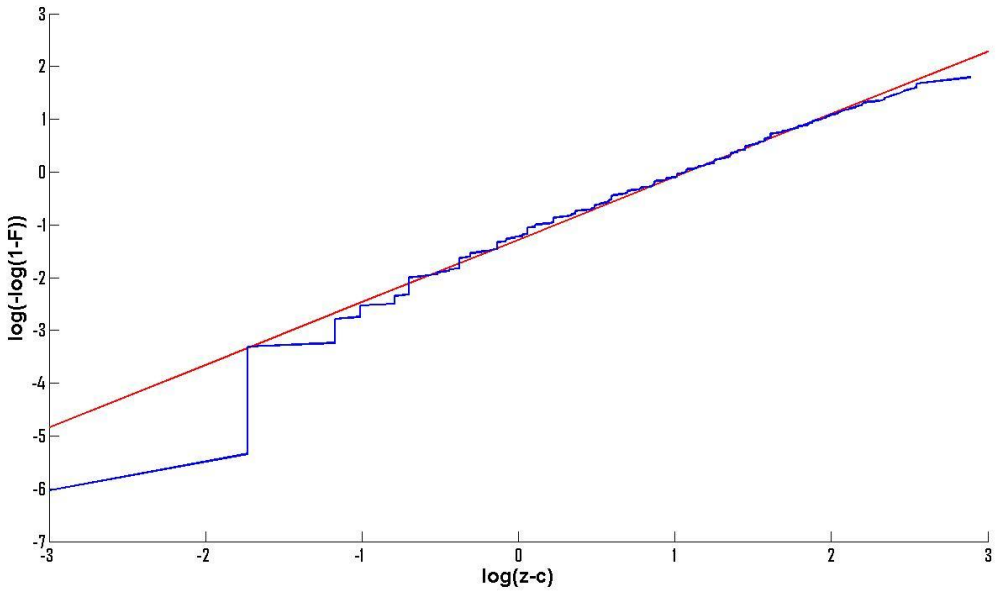
5.1. Resultater ved Tromans og Vanderschurens metode

For å se om resultatene er avhengig av stormgrensen, $H_{S,storm}$, er det beregnet ekstremverdier for tre ulike grenser, $H_{S,storm}$ = 6, 8 og 10m. Fremgangsmåten nedenfor er vist med $H_{S,storm}$ = 8m, men er helt tilsvarende for $H_{S,storm}$ = 6 og 10m.

Vi finner først fordelingen til Z. Det er foreslått (Haver og Bergstrøm (2009)) å tilpasse en 3-parameter Weibullfordeling til Z.

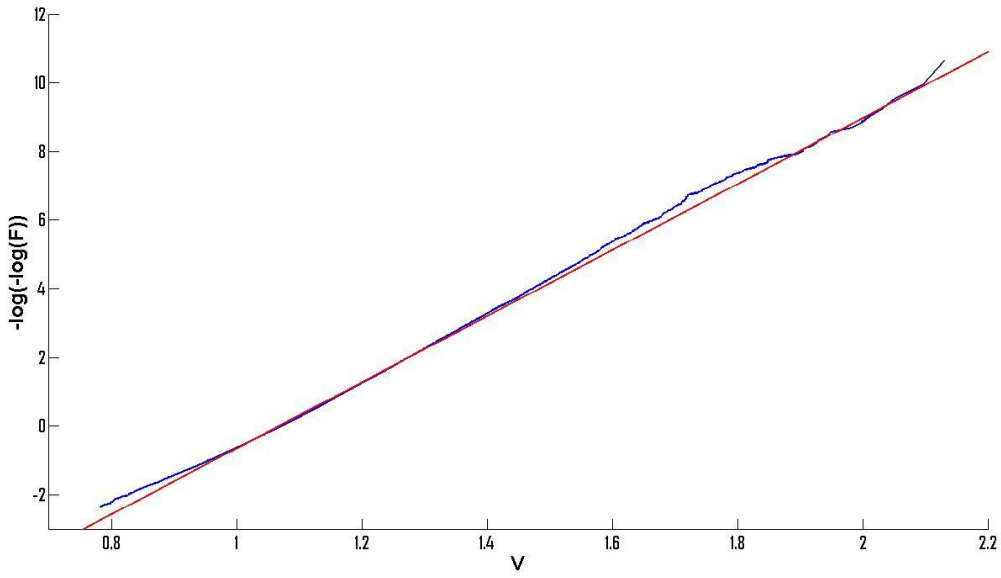
$$F_Z(z) = 1 - \exp\left(-\left(\frac{z-c}{b}\right)^a\right) \tag{27}$$

For å finne verdien til c i ligning (27) er det prøvd ut med forskjellige verdier inntil linjen i et Weibullpapir blir rett. De øvrige parametrene, a og b, er funnet i Matlab ved hjelp av funksjonen *wblfit* som gir sannsynlighetsmaksimeringsestimatene (ML-estimater) til a og b ut fra de observerte verdiene. Grunnen til at det her er benyttet ML-estimater for Weibullparametrene er at disse er anbefalt når man har et stort antall observasjoner (Amari (2008)). I dette tilfellet har vi minimum 99 (når $H_{S,storm}$ = 10m) som er antatt å være stort nok. Å benytte ML-estimatene er og mye enklere fordi disse enkelt kan finnes i Matlab ved hjelp av kommandoen wblfit. Verdiene på parametrene er funnet til: a = 1,19, b = 2,93 og c = 14,77. I figur 17 er Z plottet i et Weibullpapir. Tilpasset fordeling er plottet som rød linje, observasjonene av Z som blå linje.



Figur 17: Observerte Z-verdier plottet i et Weibullpapir (blå linje) for $H_{S,storm}$ = 8m. Rød linje er tilpasset fordeling, a = 1,19, b = 2,93 og c = 14,77 (i ligning (27)).

Fordelingen til forholdet, V, finnes ved å simulere mange tidsserier av X, bruke alle forholdene V, og tilpasse en Gumbelfordeling til disse. Verdier på parametrene er estimert ved momentmetoden, som i kapittel 3.1., og er: α_V = 1,07 og β_V = 0,104. De observerte forholdene, V, er plottet som blå linje i et Gumbelpapir i figur 18, tilpasset fordeling er vist som rød linje.



Figur 18: Forholdet, V, plottet for hver storm ($H_{S,storm}$ = 8m) i et Gumbelpapir. Rød linje er tilpasset fordeling, α_V = 1,07 og β_V = 0,104 i ligning (23).

Fra de ulike simulasjonene av X er det observert at fordelingen til V er veldig lik for de ulike stormgrensene når Ψ = 1. V følger også fint en Gumbelfordeling for de ulike stormgrensene når Ψ =

1,5 og Ψ = 2, men da med andre verdier for parametrene α_V og β_V. Fordelingen til Z må imidlertid finnes separat for hver av stormgrense og for hver Ψ-verdi.

Integralet i ligning (24) løses numerisk i Matlab for ulike returperioder og stormgrenser. Resultatene er gitt i tabell 2 nedenfor.

	$H_{S,storm}$ = 6m (1165)	$H_{S,storm}$ = 8m (416)	$H_{S,storm}$ = 10m (99)
X_{100}	38,5	37,2	37,6
X_{1000}	≈45	44	45
$X_{10\,000}$	≈55	≈52	53,5

Tabell 2: Ekstremverdier for ulike stormgrenser. Tallene i parentes bak stormgrensen angir antall stormer i tidsserien av H_S og T_P man har observert med denne stormgrensen. Resultatene gjelder for Ψ = 1.

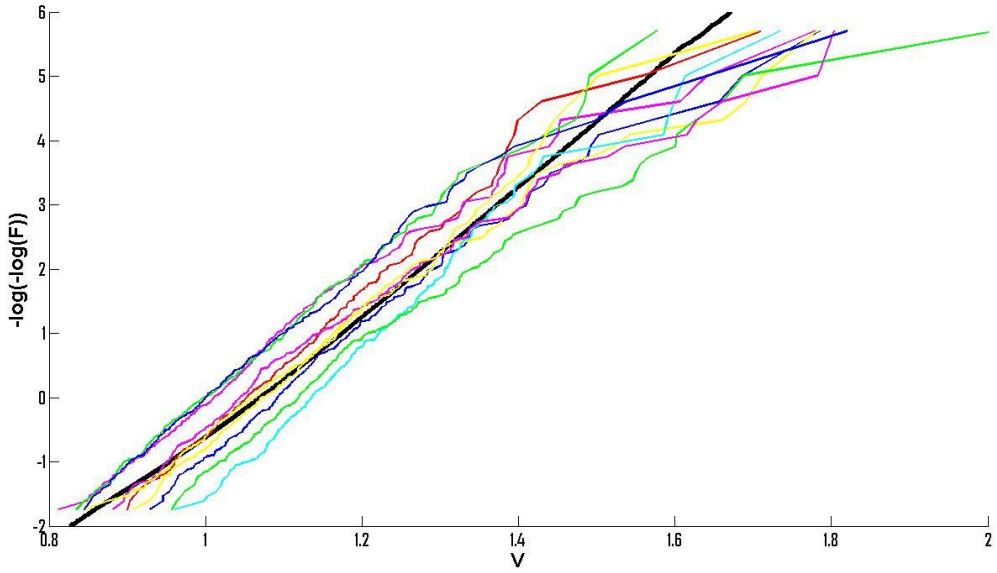
Grunnen til "≈" tegnet i tabell 2 er numeriske unøyaktigheter ved integralet utført i ligning (24). Den numeriske integrasjonen har en feil i størrelsesorden 10^{-7}, mens sannsynlighetene beregnet der "≈" tegnet er, er av størrelsesorden 10^{-5} til 10^{-6}. Resultatene må derfor sees på som omtrentlige.

Man kan se av tabell 2 at stormgrensen ikke har veldig stor betydning. Variasjonene i resultatene kan like gjerne skyldes tilnærminger i tilpasning av fordelinger eller numeriske unøyaktigheter.

5.2. Vurdering av med metoden

Som med metoden beskrevet i kapittel 2 er også denne metoden avhengig av numerisk integrasjon med høy nøyaktighet når man skal beregne ekstremverdiene. Dette gjør at resultatene kan bli unøyaktige hvis ikke den numeriske løsningsmetoden er nøyaktig nok. Men etter hvert som datakapasiteten generelt blir bedre og bedre vil nok dette problemet bli mindre og mindre.

Hvor god antagelsen om at forholdet V = Y/Z er uavhengig av stormintensitet kan man vurdere ved å simulere n tidsserier av X, deretter velge seg noen tilfeldige stormer, og plotte de n forholdene, V, for hver av disse utvalgte stormene i et Gumbelpapir. Hvis fordelingen til V er uavhengig av stormintensiteten vil plottene til disse ulike stormene bli noenlunde like. Hvis derimot forholdet V er avhengig av stormintensiteten vil disse plottene bli ulik gitt at de utvalgte stormene er ulik. I figur 19 nedenfor er ti tilfeldige stormer valgt, 200 simulasjoner av X er gjennomført, og de 200 forholdene, V, for hver av disse ti stormene er plottet i et Gumbelpapir (fargede linjer). Den sorte linjen er forholdene, V, når kun et forhold fra alle de ulike stormene er tatt med, denne er utgangspunktet for fordelingen tilpasset V gitt i ligning (23).



Figur 19: Forholdene, V, for ti ulike stormer (fargede linjer) og for alle stormene (sort linje).

Av figur 19 ser man at linjene for de ulike stormene ikke er helt lik. Man kan imidlertid se at stigningstallene er ganske like. Fordi dette er plottet i et Gumbelpapir, betyr det at skalaparameteren β_V i ligning (23) er ganske lik uansett storm, ergo, uavhengig av stormintensitet. På grunn av forskyvningen i linjene vil lokasjonsparameteren α_V være noe ulik i de ulike stormene, ergo, α_V er ikke uavhengig av stormintensitet. Forskjellen er imidlertid ikke veldig stor, antagelsen om at fordelingen til V er uavhengig av stormintensitet er derfor "grei nok". For å finne den endelige fordeling til V er det benyttet et forhold fra hver storm, tilsvarende den sorte linjen i figur 19. Som i kapittel 3 og 4 er mange tidsserier av X simulert, og alle forholdene fra denne "lange" tidsserien er benyttet slik at informasjonen i tidsserien av H_S og T_P benyttes maksimalt.

Det har vist seg at denne metoden er ganske følsom for hvilken fordeling man benytter for Z. Plotter man de observerte verdiene for Z i ulike sannsynlighetspapir ser man at det er flere fordelinger som kan passe bra til de observerte dataene. I et Paretopapir ligger punktene også på en fin linje, særlig for $H_{S,storm}$ = 8 og 10m og Ψ = 1,5 og 2 (se figurer i appendiks D). Hvis man tilpasser en Paretofordeling til Z vil man allikevel få helt andre resultater enn hvis man benytter en 3-parameter Weibullfordeling. For eksempel, når Ψ = 2, finner man $X_{10\,000}$ til å bli 1210 når Z er modellert med en Weibullfordeling. Tilpasser man i stedet en Paretofordeling finner man $X_{10\,000}$ = 1760, noe som utgjør en forskjell på ca 45 %. Fordi det er anbefalt (Haver og Bergstrøm (2009)) å benytte en 3-parameter Weibullfordeling for å modellere Z, er denne fordelingen benyttet. Resultatene blir da også mer i samsvar med resultatene fra de andre metodene. Grunnen til de store forskjellene er trolig den tykke halen i Paretofordelingen sammenlignet med Weibullfordelingen.

6. Konturlinjer

For komplekse problemer kan det være vanskelig å finne fordelingen til X i en sjøtilstand, og modelltesting eller simulering kan være nødvendig. Hvis man skulle gjort dette for alle mulige sjøtilstander ville det beregningsmessig blitt svært kostbart. En bedre måte er derfor å finne hvilke sjøtilstander, dvs. kombinasjoner av H_S og T_P, man kan regne med å få iløpet av for eksempel 100 år, velge den verste ut av disse sjøtilstandene, og deretter kjøre en modelltest / simulasjon i denne ene sjøtilstanden. For å finne denne verste sjøtilstanden brukes såkalte konturlinjer.

Fordi ekstremverdiberegninger er enkelt hvis man kun har en sjøtilstand å ta hensyn til, brukes også konturlinjer selv om man kjenner fordelingen til X og kunne ha benyttet noen av metodene nevnt i de forrige kapitlene. Det vil derfor bli sett på hvordan resultatene fra en slik analyse blir sammenlignet med de andre metodene som har vært presentert. Konturlinjer er også nyttige i startfasen til et prosjekt for å få oversikt over hvilke sjøtilstander konstruksjonen bør dimensjoneres for, og hvilke sjøtilstander man kan forvente å oppleve iløpet av levetiden.

6.1. Ulike konturlinjer

En konturlinje er en linje der alle punktene langs linjen gir kombinasjoner av H_S og T_P. Det er flere måter å lage konturlinjer på. Metoder som har vært benyttet går ut på å lage en linje slik at enten alle punktene på linjen tilsvarer en gitt sannsynlighet for overskridelse (FORM-linjer), eller hvor sannsynligheten for å havne utenfor linjen er en gitt sannsynlighet (isolinjer). Disse ulike metodene vil bli diskutert under.

6.1.1. Isolinjer

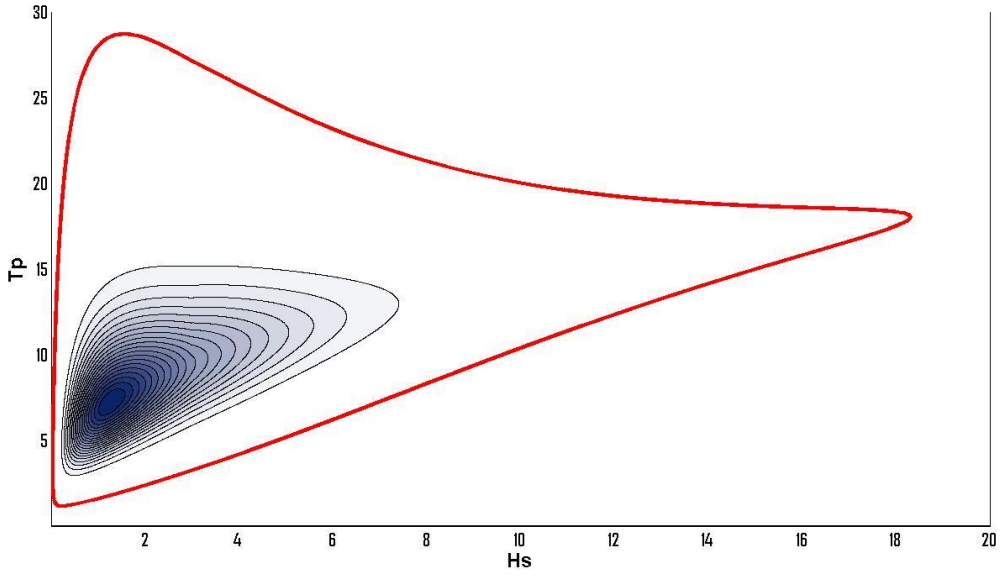
En isolinje er en linje, i domenet til H_S og T_P, der sjansen for å havne utenfor denne linjen er en gitt sannsynlighet, for eksempel $1/N_{100}$. For at denne linjen skal være entydig definert er det også et krav at linjen ligger på konstant sannsynlighetstetthet, det vil si, $f(h_S,t_P)$ er konstant langs denne linjen. En 100-års isolinje er derfor definert ved:

$$\iint\limits_{\{H_S,T_P\}\subset Isolinjen} f_{H_S,T_P}\left(h_S,t_P\right)dh_S dt_P = 1-\frac{1}{N_{100}} \tag{28}$$

$$f_{H_S,T_P}\left(h_S,t_P\right) = \text{konstant langs linjen} \tag{29}$$

Ligning (28) og (29) definerer en isolinje hvor gjennomsnittlig en hendelse per 100 år vil havne utenfor denne linjen. Den verste kombinasjonen av H_S og T_P langs denne linjen benyttes i fordelingen til X gitt i ligning (1). En passende kvantil i denne fordelingen gir så 100-års verdien for X. Vanligvis har en 0,9 kvantil vært benyttet. Hvor riktig dette er imidlertid usikkert, og vil variere fra problem til problem.

Figur 20 nedenfor viser 100-års isolinje for fordelingen funnet i kapittel 2.2.



Figur 20: 100-års isolinje (rød linje).

Ulemper med isolinjemetoden er at den ikke forteller noe om sannsynligheten for å overskride ulike punkt langs linjen. Hvis fordelingen er veldig usymmetrisk kan forskjellene være store, dette er forsøkt vist i figur 21 nedenfor. Her er et snitt av en usymmetrisk tetthetsfunksjon, f(x), tegnet inn sammen med en isolinje. Man ser at det er mye større sjanse for å havne til høyre for isolinjen enn til venstre.

Figur 21: Sannsynligheter for overskridelse av isolinje

100-års isolinjen vil i alle praktiske situasjoner overskride 100-års verdiene (=verdiene som overskrides gjennomsnittlig en gang iløpet av 100 år) til H_S og T_P man ville fått fra marginalfordelingene. Dette er fordi 100-års verdien til H_S også kan sees på som en konturlinje der sjansen er $1/N_{100}$ for å havne til høyre for denne linjen, dvs. få en større H_S verdi enn 100-års verdien. Sjansen for å havne utenfor 100-års isolinjen er også $1/N_{100}$, men i dette tilfellet er ikke hele området til venstre for 100-års verdien til H_S tatt med, isolinjen må derfor overskride 100-års linjen til H_S for å inneholde et stort nok volum. Tilsvarende resonnement gjelder for 100-års verdien for T_P, og for 1000 og 10 000 års isolinjer. Hvor mye 100-års isolinjen vil overskride 100-års verdiene fra marginalfordelingene avhenger av den simultane fordelingen til H_S og T_P.

Hvis man velger det verste punktet langs 100-års isolinjen som dimensjonerende verdipar for H_S og T_P vil man alltid få et konservativt estimat. Dette fordi sannsynligheten for å havne utenfor isolinjen er

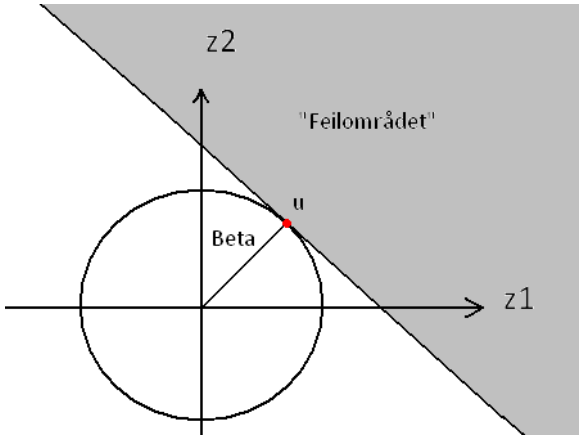
$1/N_{100}$, men det er ikke sikkert at alle områdene utenfor isolinjen er kritiske. Å havne utenfor isolinjen, og samtidig havne i et kritisk område vil derfor alltid vil være mindre eller lik $1/N_{100}$. Å benytte isolinjer vil derfor være en trygg metode.

6.1.2. FORM-linjer

FORM (first-order reliability method) metoden går ut på at man benytter Rosenblatt-transformasjonen og transformerer variablene, H_S og T_P, til to uavhengige standard normalfordelte variabler, Z_1 og Z_2. Rosenblatt transformasjonene er gitt ved:

$$\begin{aligned} F_{H_S}\left(h_S\right) &= \Phi\left(z_1\right) \\ F_{T_P|H_S}\left(t_P \mid h_S\right) &= \Phi\left(z_2\right) \end{aligned} \tag{30}$$

Her er ϕ er den kumulative fordelingen til en standard normalfordelt variabel. Prinsippet med FORM er at man i normal-rommet, dvs. med variablene Z_1 og Z_2, kan tilnærme feiloverflaten som er rett linje. Feiloverflaten vil si grensen til det området, i domenet til H_S og T_P, der man ikke ønsker å være. Dette er vist som grått området i figur 22. Under denne antagelsen er sannsynligheten for å havne i dette feildomenet lik ϕ(-β) hvor β er korteste avstand fra sentrum til den rette linjen. Punktet der feilflaten er nærmest kalles designpunktet, u. β er altså avstanden fra origo til designpunktet. Dette er forsøkt vist på figur 22 nedenfor.

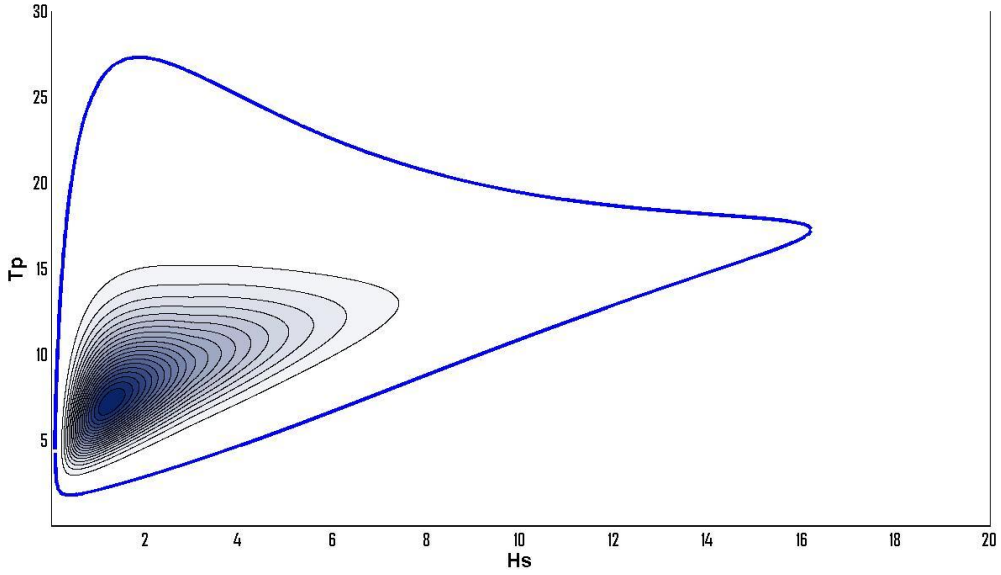


Figur 22: Prinsippskisse for FORM metoden

Dette betyr at alle punktene langs sirkelen i figur 22 vil med samme sannsynlighet overskrides ut i et feilområde, gitt at feilområdet er en rett linje i normal-rommet. Denne sirkelen kan derfor transformeres tilbake til H_S-T_P-rommet og brukes som en konturlinje der alle punktene langs linjen vil bli overskredet ut i feilområdet med samme sannsynlighet, ϕ(-β). Hvis man vil finne en 100-års konturlinje finner man først β slik:

$$\Phi\left(\beta_{100}\right) = 1 - \frac{1}{N_{100}} \rightarrow \beta_{100} = \Phi^{-1}\left(1 - \frac{1}{N_{100}}\right) = 4,4985 \tag{31}$$

Sirkelen i normal-rommet definert ved $z_1^2 + z_2^2 = \beta_{100}^2$ kan derfor transformeres tilbake til H_S-T_P-rommet og benyttes som 100-års konturlinje. Dette er gjort og resultatet er vist i figur 23. For mer om FORM tilnærmingen, se for eksempel (Madsen et. al (1986)).



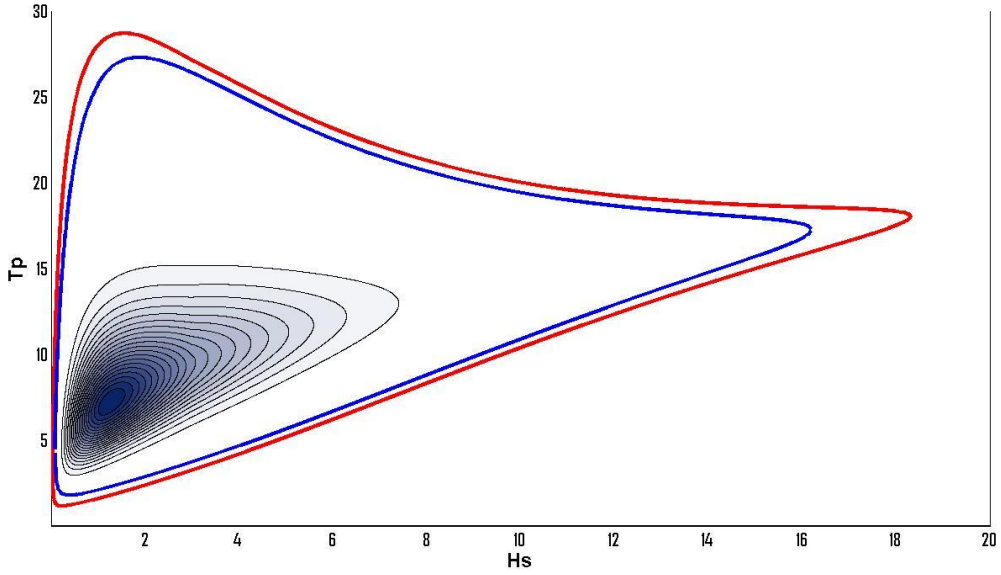
Figur 23: 100-års FORM-linje.

En usikkerhet ved FORM metoden er antagelsen om at feilflaten kan tilnærmes som en rett linje i normal-rommet. Feilen er imidlertid ofte liten fordi mesteparten av volumet som bidrar til sannsynligheten vil ligge rett utenfor designpunktet, u, og for dette lille området er tilnærmingen ved en rett linje ok. Dette er diskutert mer i (Madsen et. al (1986)).

En annen ting man må være observant på er at man vil få to ulike konturlinjer hvis man enten bruker $f(h_S)$ og $f(t_P|h_S)$, eller $f(t_P)$ og $f(h_S|t_P)$ i transformasjonen gitt i ligning (30), se appendiks B for et eksempel. Dette kan enkelt forklares fordi; hvis man definerer transformasjonen som i ligning (30), og bruker β_{100}, vil man få en konturlinje som går fra $H_{S,min100}$ til $H_{S,max100}$ hvor $H_{S,min100}$ og $H_{S,max100}$ er henholdsvis den verdien av H_S som underskrides og overskrides gjennomsnittlig én gang iløpet av 100 år gitt ved marginalfordelingen til H_S. Mellom disse to ytterpunktene vil konturlinjen ta T_P verdier i samsvar med siste ligning i ligning (30). Men det er ingenting som tilsier at den største av disse T_P verdiene vil gå gjennom $T_{P,max100}$, noe man vill fått hvis man hadde definert transformasjonen ved $f(t_P)$ og $f(h_S|t_P)$ i stedet for $f(h_S)$ og $f(t_P|h_S)$. Derfor vil de to konturlinjene være ulik. Ulikheten vil variere etter hvor lik de to fordelingene (til H_S og T_P) er normalfordelt; jo likere de er, jo mindre vil forskjellen være. Forskjellen vil også minke jo større β er (Haver og Winterstein (2010)).

Det vanlig å definere transformasjonen slik at marginalfordelingen til den viktigste variabelen benyttes, også betinge den andre variabelen på denne (Haver og Winterstein (2010)). Da er man sikker på at maksimalverdiene til konturlinjen for den viktigste variabelen er de samme som man ville fått ved å beregne disse direkte fra marginalfordelingen. Mellom disse ytterpunktene gir konturlinjen en god indikasjon på hvordan den andre variabelen vil variere. I vårt tilfelle er det antatt at H_S er den viktigste variabelen, transformasjonen gitt i ligning (30) er derfor benyttet for å finne konturlinjen.

Figur 24 viser 100-års isolinjen (rød) og 100-års FORM-linjen (blå) for fordelingen til H_S og T_P funnet i kapittel 2.2.



Figur 24: 100-års isolinje (rød) og 100-års FORM-linje (blå).

Man kan legge merke til at høyeste verdi for H_S langs FORM-linjen tilsvarer 100-års verdien for H_S man ville fått fra marginalfordelingen gitt i ligning (6).

Man må huske at isolinjen og FORM-linjen er bygget på helt forskjellige prinsipper, og kan derfor strengt tatt ikke sammenlignes. Isolinjen må tolkes på den måten at man vil havne utenfor denne gjennomsnittlig en gang per 100 år, dette er ikke tilfelle med FORM-linjen. Sannsynligheten for å havne utenfor FORM-linjen kan finnes ved å beregne volumet innenfor denne linjen. Dette volumet finnes enklest ved integrasjon i normal-rommet. Sannsynligheten for å havne utenfor er da 1 – volumet innenfor linjen. For 100-års FORM-linjen har vi:

$$V_{FORM} = 2\pi \int_0^{\beta_{100}} x \cdot f_Z(x)dx = 2\pi \int_0^{\beta_{100}} \frac{x}{2\pi} e^{\left(-\frac{x^2}{2}\right)} dx = 1 - e^{-\frac{\beta_{100}^2}{2}} = 0,9999597 \quad (32)$$

$$P(\text{havne utenfor FORM-linjen}) = 1 - V_{FORM} = 4,03 \cdot 10^{-5} \quad (33)$$

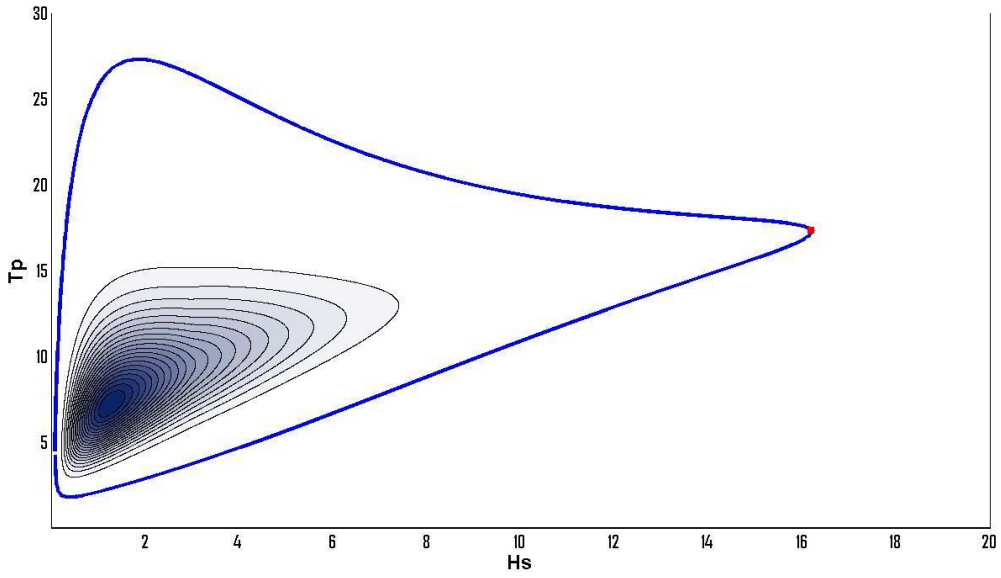
Til sammenligning er sannsynligheten for å havne utenfor 100-års isolinjen:

$$P(\text{havne utenfor iso-linjen}) = \frac{1}{N_{100}} = 3,43 \cdot 10^{-6} \quad (34)$$

Dette viser at sannsynligheten for å overskride FORM-linjen er over 10 ganger større i dette tilfellet. Punktene langs FORM-linjen må derfor tolkes som punkt som gjennomsnittlig vil overskrides en gang per 100 år, og med overskrides menes å havne i et feilområde som kan tilnærmes med en rett linje i normal-rommet jfr. figur 22.

6.2. Resultater fra konturlinjene

Fordi FORM-linjene er de som er mest benyttet idag er verste sjøtilstand funnet ved hjelp av denne konturlinjen. 100-års sjøtilstanden som gir størst 0,9 kvantil verdi for X, gitt ved fordelingen til X i ligning (1), er sjøtilstanden merket med rød prikk langs 100-års FORM-konturlinjen i figur 25.



Figur 25: Verste sjøtilstand (rød prikk).

Med 0,9 kvantilen i utvalgte sjøtilstand finner man at:

$X_{100} = 38{,}33$

Ved hjelp av 1000 og 10 000-års FORM-linjene finner man tilsvarende:

$X_{1000} = 43{,}35$

$X_{10\,000} = 48{,}17$

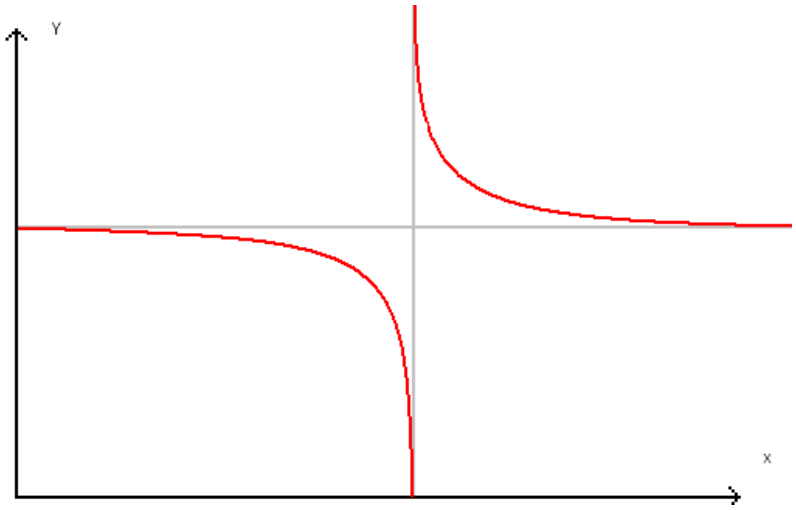
Dette er resultatene for $\Psi = 1$. Se tabell 4 og 5 for resultater når $\Psi = 1{,}5$ og $\Psi = 2$.

Det er interessant å se at resultatene for 100 og 1000-års verdiene for X er i samsvar med de øvrige resultatene, men 10 000-års verdien er noe lavere. Hva dette skyldes er usikkert. Hvis man isteden for å bruke 0,9 kvantilen bruker 0,95 kvantilen for 10 000-års verdien finne man $X_{10\,000}$ til å bli 51 som er mer i samsvar med de andre resultatene.

6.3. Alternative konturlinjer

En ny type konturlinjer er foreslått (Salvadori et. al (2007)) for å predikere ekstremverdier for kombinasjoner av flere variabler. Metoden benytter kopulaer (se appendiks F) for å lage ”eller–linjer” og ”og–linjer”. For et bivariat tilfelle med variablene x og y, gir en ”eller-linje” alle kombinasjonene av

x og y der en eller begge variablene er mindre enn verdiene på linjen med sannsynligheten q. Tilsvarende gir en ”og-linje” alle kombinasjonene av x og y hvor begge variablene er mindre enn verdiene på linjen med sannsynligheten q, se figur 26.



Figur 26: Eksempel på de ”nye” konturlinjene. Rød linje oppe til høyre er en ”og-linje”, mens den røde linjen nede til venstre er en ”eller-linje”. P(X<x og Y<y) er konstant lik q langs ”og-linjen”, mens P(X>x og Y>y) er konstant lik 1-q langs ”eller-linjen”.

Som for iso og FORM-linjene er poenget også med disse linjene å finne den verste kombinasjonen av x og y, og bruke disse verdiene som dimensjonerende kombinasjonspar. Dette er imidlertid ikke alltid så enkelt, spesielt ikke for ”og-linjene” som går til uendelig. Se (Salvadori et. al (2007)) for en bedre diskusjon.

Disse ”og” og ”eller-linjene” kunne også ha vært funnet ved hjelp av den simultane fordelingen til x og y, men ved å bruke Arkimediske kopulaer blir det mye enklere. Dette fordi man da har et analytisk utrykk for nivålinjene, det vil si linjer der P(X<x og Y<y) og P(X>x og Y>y) er konstant.

Disse ”nye” konturlinjene kan være nyttig i tilfeller der store verdier av både x og y er kritisk. Når man har en dynamisk responsstørrelse som er avhengig av bølgehøyde og bølgeperiode er ikke dette nødvendigvis tilfellet. I slike tilfeller er det oftest periodestørrelser nær egenperioden som er kritisk. Konturlinjene presentert i dette avsnittet er derfor ikke så godt egnet for slike problemer. Men for situasjoner der store verdier for begge variablene er kritisk kan metoden være interessant å benytte. Dette kan for eksempel være responsstørrelser avhengig av bølgehøyde og vind.

Bivariate og multivariate ekstremverdifordelinger er og et fagfelt som er utviklet i det siste. Her er ideen å modellere vektorvise ekstremer (Salvadori et. al (2007)), altså ekstremverdier for hver av variablene. Men metoden tar ikke hensyn til om maksimaene til de ulike variablene opptrer samtidig. Denne metoden er derfor ikke til stor nytte for situasjoner som er avhengig av at ”ugunstige” kombinasjoner av begge variablene skal opptre samtidig.

7. Resultat

For enkelhetsskyld vil de ulike metodene diskutert i de foregående kapitlene heretter bli kalt:

Metode 1 = Klassisk langtidsanalyse, kapittel 2
Metode 2 = Gumbelfordeling til årlige maksima, kapittel 3
Metode 3 = Næss-Gaidai-metoden, kapittel 4
Metode 4 = Tromans og Vanderschurens metode, kapittel 5
Metode 5 = Konturlinjemetoden (FORM), kapittel 6

7.1. Resultattabell

100, 1000 og 10 000 års verdiene for X beregnet ved de forskjellige metodene, og for de tre ulike verdiene av Ψ er samlet i tabell 3, 4 og 5 nedenfor. For metode 4 er $H_{s,storm}$ = 10m benyttet.

	Metode 1	Metode 2	Metode 3	Metode 4	Metode 5
X_{100}	38,0	37,8	37,7	37,6	38,3
X_{1000}	44,6	44,8	44,3	45,0	43,4
$X_{10\,000}$	51,4	51,8	50,9	53,5	48,2

Tabell 3: Ψ = 1

	Metode 1	Metode 2	Metode 3	Metode 4	Metode 5
X_{100}	162	158	159	160	169
X_{1000}	203	195	203	205	202
$X_{10\,000}$	247	233	254	252	236

Tabell 4: Ψ = 1,5

	Metode 1	Metode 2	Metode 3	Metode 4	Metode 5
X_{100}	684	644	675	675	735
X_{1000}	908	825	896	933	937
$X_{10\,000}$	1159	1006	1144	1210	1152

Tabell 5: Ψ = 2

7.2. Kommentarer til resultatene

Med tanke på unøyaktigheter i modelltilpasninger og avrundinger i beregninger kan man si at resultatene grovt sett er ganske like. Noen resultater skiller seg allikevel litt ut.

For metode 1 og 3 er resultatene for de ulike verdiene av Ψ forholdsvis like. Resultatene ved metode 2 er og ganske lik for Ψ = 1, og Ψ = 1,5. For Ψ = 2 derimot, er resultatene betydelig lavere enn ved de andre metodene. Det er ganske sikkert at det er metode 2 som ikke er helt god her. For Ψ = 2 følger årlige maksima ganske dårlig en Gumbelfordeling, se figur C i appendiks C. Resultatene ved metode 4 er og ganske lik resultatene fra metode 1 og 3, men sammenligner man 10 000-års verdiene virker de

å være noe høy ved denne metoden, særlig for Ψ = 2. Metode 5 predikerer litt høye 100-års verdier sammenlignet med de øvrige metodene.

Ved metode 1 benytter man fordelingen til H_S og T_P, og i metode 4 modelleres stormene. Man kan derfor bevege seg utenfor de observerte verdiene av disse variablene, og ta høyde for "uobserverte stormer". Forskjellen i resultatene sammenlignet med metode 3, som ikke tar med effekten av "uobserverte stormer", er imidlertid ubetydelig. Dette er trolig fordi med en såpass lang tidsserie med verdier av H_S og T_P vil denne effekten ikke ha så stor betydning.

Ved å tilpasse en fordeling til H_S og T_P som gjort i kapittel 2.2. antar man at hver observasjon / sjøtilstand er uavhengig. Dette er en konservativ antagelse (se kapittel 2.4.), men som man ser av tabell 3, 4 og 5 er ikke resultatene betydelig høyere ved metoden 1 enn ved de andre metodene.

8. Konklusjon og videre arbeid

Av tabell 3, 4 og 5 ser man at resultatene ved de ulike metodene grovt sett er ganske like. Hvilke resultater som er nærmest de virkelige ekstremverdiene er umulig å si i og med at man ikke vet hvilken statistisk modell naturen har valgt for H_S og T_P. Hvilken metode som er "best" må derfor vurderes ut fra hvor god de ulike antagelsene bak hver av metodene er, hvor godt de ulike metodene benytter informasjonen i de tilgjengelige dataene, og hvor enkel de er å bruke. En vurdering av de ulike metodene er oppsummert nedenfor:

Metode:	**Positivt:**	**Negativt:**
Metode 1: Klassisk langtidsanalyse	- Benytter all informasjonen i de tilgjengelige dataene ved å bruke alle observasjonene av H_S og T_P - Ved å finne en parametrisk fordeling til H_S og T_P tar man høyde for "uobserverte stormer"	- Beregningskrevende integraler som må løses numerisk - Antar uavhengighet mellom sjøtilstandene, noe som ikke stemmer med virkeligheten. Dette er imidlertid en konservativ antagelse.
Metode 2: Gumbelfordeling til årlige maksima	- Veldig enkel å benytte, og lite beregningskrevende - Ved å benytte årlige maksima slipper man problemer med klumping	- Bruker veldig lite av de tilgjenglige dataene - Det er ikke alltid årlige maksima vil følge en Gumbelfordeling
Metode 3: Næss-Gaidai-metoden	- Enkel å benytte, og lite beregningskrevende - Tar høyde for avhengighet mellom sjøtilstandene	- Ingen åpenbare negative sider ved denne metoden
Metode 4: Tromans og Vanderschurens metode	- Tar høyde for "uobserverte stormer"	- Tunge numeriske utregninger - Veldig følsom for valg av fordeling for Z.
Metode 5: Konturlinjemetoden (FORM)	- Enkel å benytte - Gir og en oversikt over hvilke sjøtilstand man kan forvente å få iløpet av aktuelle returperiode	- Vanskelig å si hvor god antagelsen om at 0,9 kvantilen i fordelingen til X (i verste sjøtilstand) kan brukes som ekstremverdi uten å sammenligne med mer nøyaktige metoder.

Resultatene ved metode 1 og 3 er de man har størst grunn for å tro på. Gumbelantagelsen i metode 2 er ikke alltid helt god (se kapittel 3.1.3. og appendiks C), metode 4 virker å være veldig følsom for valg av fordeling til Z (se kapittel 5.2. og appendiks D), og det er vanskelig å si hvor god 0,9 kvantil antagelsen ved metode 5 er uten å sammenligne med andre metoder.

Metode 1 og 3 bygger også på antagelser, men disse er mindre usikker enn ved de andre metodene. Antagelsen om at sjøtilstandene er uavhengig slik som i metode 1 gir en feil, men feilen man innfører her er veldig liten, trolig i størrelsesorden noen få prosent (Haver (2010)). Metode 3 bygger på Poissonantagelsen, at krysninger over et høyt nivå vil være tilnærmet Poissonfordelt. Denne antagelsen er og antatt å være god som diskutert i kapittel 4.2.

Av metode 1 og 3 er metode 3 betydelig enklere å bruke enn metode 1, og den krever mye mindre tunge numeriske utregninger. Man kan derfor konkludere med at metode 3 er den beste og enkleste av metodene det her er sett på for å predikere ekstremverdier.

I denne oppgaven har det vært fokusert på å forstå og sammenligne eksisterende metoder. I det videre arbeid ville det vært interessant og sett på hva som eventuelt kan gjøres for å forbedre metodene som brukes, og om det kan være andre måter å predikere ekstremverdier på enn de som benyttes idag.

Symbolliste

a	parameter i formel for nivåoppkrysningsfrekvens
a	parameter i fordeling til Z
b	parameter i formel for nivåoppkrysningsfrekvens
b	parameter i fordeling til Z
c	parameter i formel for nivåoppkrysningsfrekvens
c	parameter i fordeling til Z
C()	kopulafunksjonen
$e_n()$	oppkrysningsfrekvensen betinget på de n foregående observasjonene
exp()	eksponentialfunksjonen
f()	fordelingsfunksjon (tetthetsfunksjon)
F()	kumulativ fordelingsfunksjon
H_S	signifikant bølgehøyde
$H_{S,storm}$	H_S nivå som skiller storm fra ikke-storm
$H_{S,min100}$	H_S verdi som underskrides gjennomsnittlig en gang per 100 år
$H_{S,max100}$	H_S verdi som overskrides gjennomsnittlig en gang per 100 år
k	parameter i fordeling til H_S
log()	den naturlige logaritme
m	meter
N	antall observerte maksima (=antall 3-timers perioder i tidsserien for H_S og T_P)
N_X	antall maksima iløpet av x år
N_{storm}	antall stormer (perioder der H_S>$H_{S,storm}$)
$N_{storm,X}$	antall stormer iløpet av x år
n	antall / telleparameter
p	parameter i fordelingen til X
q	returperiode i år
q()	funksjon i formel for nivåoppkrysningsfrekvens
r	parameter i fordelingen til X
R^2	koeffisient for regresjonsmodell
s	sekund
t_0	egenperiode
T_P	topp periode
T_X	antall tidssteg iløpet av x år
$T_{P,min100}$	T_P verdi som underskrides gjennomsnittlig en gang per 100 år
$T_{P,max100}$	T_P verdi som overskrides gjennomsnittlig en gang per 100 år
u	designpunkt ved FORM-metoden
V	volum
V	forholdet mellom Y og Z, V = Y/Z
X	responsstørrelsen
X(t)	responsprosessen (responsstørrelsen X ved tidssteg t)
x	nivå/verdi for X, hjelpevariabel
Y	største observerte responsverdi i hver storm, dvs. største X-verdi iløpet av en storm
y	hjelpevariabel
Z	største forventede responsverdi i hver storm, dvs. største α-verdi iløpet av en storm

z_1 standard normalfordelt variabel nr 1

z_2 standard normalfordelt variabel nr 2

α variabel i Gumbelfordelingen

β variabel i Gumbelfordelingen

β_X radius i ”normal-rommet” for x-års FORM-linje

μ forventningsverdi

λ parameter i fordeling til H_S

λ parameter i fordelingen til X

$\nu^+()$ nivåoppkrysningsfrekvens

σ^2 varians

$\phi()$ standard kumulativ normalfordeling

Ψ parameter i fordelingen til X

Referanser

Amari, S. (2008): “Weibull parameter estimation methods”. Reliability Articles, Relex.

Gut, A. (2005): “Probability: A Graduate Course”. Springer.

Haver, S. (2009): Forelesningsnotater fra kurset TMR4195 våren 2009.

Haver, S. (2010): Personlig kommunikasjon.

Haver, S., og Bergsvik, J.E. (2009): “Extreme response in a hurricane governed offshore region: Uncertainties related to limited amount of data and choice of method of prediction”.

Haver, S., og S. Winterstein (2010): Personlig kommunikasjon.

Leira, B. J. (2005): ”TMR4235 Stochastic Theory of sealoads. Probabilistic modelling and estimation”, Fakultet for Ingeniørvitenskap og teknologi, NTNU Trondheim.

Løvås, G. G. (2004): “Statistikk for universiteter og høyskoler”, Universitetsforlaget.

Madsen, H., S. Krenk og N. Lind (1986): ”Structural Safety”, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1986.

Myrhaug, D. (2005): ”TMR4235 Stochastic Theory of sealoads. Statistics of narrow band processes and equivalent linearization”, Fakultet for Ingeniørvitenskap og teknologi, NTNU Trondheim.

Nelsen, R. B. (2006): ”An introduction to copulas”, second edition. Springer.

Næss, A. & O. Gaidai (2008): “Estimation of extremes values from sampled time series”. Structural safety , 31 side 325 – 334.

Næss, A. C. T. Stansberg, O. Gaidai og R. J. Baarholm (2009): “Statistics of extreme events in airgap measurements”, Journal of Offshore Mechanics and Arctic Engineering.

Salvadori, G., C. De Michele, N. T. Kottegoda og R. Rosso (2007): “Extremes in nature, an approach using Copulas”, Springer.

Winterstein, S. (2010): Gjesteforelesning.

Åsgard Metocean Design Basis (2004): Statoil ASA.

Appendiks:

A. Bakgrunn for Gumbelantagelsen

Gumbelfordelingen er mye brukt i denne oppgaven for å modellere ekstremverdier. Grunnen til at akkurat denne fordelingen er valgt vil bli diskutert nedenfor.

Fordelingen til den største av n uavhengige identisk fordelte variabler X, er gitt ved ligning (A).

$$F_{X_{\max}}(x) = \left(F_X(x)\right)^n \tag{A}$$

Hvis n går mot uendelig får vi:

$$\lim_{n\to\infty}\left(F_X(x)\right)^n = \begin{cases} 1 & \text{hvis } F_X(x) = 1 \\ 0 & \text{hvis } F_X(x) < 1 \end{cases} \tag{B}$$

Grensefordelingen til $F_X(x)$ er derfor degenerativ. Det er imidlertid mulig å finne en grensefordeling G(x) hvis man innfører en sekvens av $\{a_n\}$ og $\{b_n > 0\}$, som en funksjon av n, slik at funksjonen

$$\lim_{n\to\infty}\left(F_X(a_n + b_n x)\right)^n = G(x) \tag{C}$$

ikke er degenerativ. Hvis ligning (C) er oppfylt tilhører F "det maksimale tiltrekningsdomenet" til G (=the maximum domain of attraction of G). Det kan vises (Gut (2005)) at, hvis det finnes en sekvens $\{a_n\}$ og $\{b_n > 0\}$ slik at ligning (C) er oppfylt, og G(x) er ikke-degenerativ, da vil G tilhøre en av tre grensefordelinger. Disse kalles, Gumbel, Weibull og Fréchet fordelinger. Hensikten med å innføre sekvensene $\{a_n\}$ og $\{b_n\}$ er å skalere variabelen slik at funksjonen ikke er degenerativ, og samtidig beholder formen til den asymptotiske fordelingen til maksimaene.

De fleste fordelinger man bruker for å modellere bølgehøyder, bølgeperioder eller responsstørrelser, som for eksempel normal, lognormal, Rayleigh, eksponential, gamma og Gumbel fordelinger, vil fordelingen til største verdi følge en Gumbelfordeling asymptotisk (Salvadori et. al (2007)). Altså, hvis F(x) tilhører en av de nevnte fordelingene over vil G(x) i ligning (C) være en Gumbelfordeling. Dette er grunnen bak antagelsen om at største verdi vil være Gumbelfordelt.

I kapittel 3 antok vi at årlige maksima fulgte en Gumbelfordeling. Men for at ligning (A) skal være riktig må de ulike observasjonene være uavhengige. Fra tidsserien med "observerte" verdien av responsstørrelsen, X, vet vi at hver av observasjonene ikke er uavhengig, Gumbelantagelsen blir derfor ikke helt riktig. Om fordelingen til X er tilhører "det maksimale tiltrekningsdomenet" til Gumbelfordelingen er heller ikke sikkert. Dessuten kan det og være at antall maksima iløpet av et år (= 2922) ikke er et stort nok antall observasjonene for at fordelingen til største verdi skal nå sin asymptotiske form.

B. Konturlinjer ved Rosenblatt transformasjonen (FORM-prinsippet)

Hensikten med dette eksempelet er å vise at en sirkel i et "normal-rom", dvs. definert ved to standard normalfordelte variabler, ikke vil gi en entydig linje når man transformerer denne sirkelen tilbake til H_S-T_P rommet ved hjelp av Rosenblatt transformasjonen. Linjen i det virkelige rommet (H_S-T_P rommet) avhenger av hvilken variabel man velger å betinge på i de to kumulative fordelingene.

Det er for enkelhets skyld antatt at begge de to variablene, H_S og T_P, er eksponentialfordelt. Dette er kun for å forenkle de matematiske utrykkene i utregningen, og er ikke realistisk iforhold til måleobservasjonene. De kumulative fordelingene til de to variablene er gitt ved:

$$\begin{aligned} F_{H_S}\left(h_S\right) &= 1-e^{-h_S} \\ F_{T_P|H_S}\left(t_P \mid h_S\right) &= 1-e^{-h_S \cdot t_P} \end{aligned} \tag{D}$$

Parameteren i fordelingen til H_S er for enkelhetsskyld satt til 1, mens parameter i fordelingen til T_P gitt H_S er satt til H_S. Dette vil si at forventningen og variansen til T_P minker når H_S øker noe som er urealistisk. Man burde heller valgt parameteren lik for eksempel en delt på H_S, men da vil ikke integralet i ligning (F) kunne løses. Man har derfor begrenset valgfrihet. Disse fordelingene vil derfor ikke være gode for å beskrive den virkelige fordelingen til H_S og T_P, men de vil kunne være realistisk i andre situasjoner.

For å finne de "motsatte" kumulative fordelingene finner vi først den bivariate tetthetsfunksjonen. Vi deriverer de to ligningene i (D) og multipliserer, dette gir:

$$f_{H_S,T_P}\left(h_S,t_P\right) = h_S \cdot e^{-h_S\left(1+t_P\right)} \tag{E}$$

Vi finner først den kumulative fordelingen til T_P:

$$f_{T_P}\left(t_P\right) = \int_0^\infty f_{H_S,T_P}\left(h_S,t_P\right)dh_S = \int_0^\infty h_S \cdot e^{-h_S\left(1+t_P\right)}dh_S = \frac{1}{\left(1+t_P\right)^2} \tag{F}$$

$$F_{T_P}\left(t_P\right) = \int_0^{t_P} f_{T_P}\left(a\right)da = \int_0^{t_P} \frac{1}{\left(1+a\right)^2}da = 1-\frac{1}{1+t_P} \tag{G}$$

Vi kan nå finne den kumulative fordelingen til H_S gitt T_P:

$$f_{H_S|T_P}\left(h_S \mid t_P\right) = \frac{f_{H_S,T_P}\left(h_S,t_P\right)}{f_{T_P}\left(t_P\right)} = \left(1+t_P\right)^2 \cdot h_S \cdot e^{-h_S\left(1+t_P\right)} \tag{H}$$

$$F_{H_S|T_P}\left(h_S \mid t_P\right) = \int_0^{h_S} f_{H_S|T_P}\left(a \mid t_P\right)da = \int_0^{h_S} \left(1+t_P\right)^2 \cdot a \cdot e^{-a\left(1+t_P\right)}da = 1-e^{-h_S\left(1+t_P\right)}\left(1+\left(1+t_P\right)h_S\right) \tag{I}$$

I ligning (G) og (I) er "a" en integrasjonsvariabel. Vi har nå to sett av kumulative fordelinger. Vi kan da definere to Rosenblatt transformasjoner, og to tilhørende konturlinjer:

Konturlinje nr. 1 er gitt ved transformasjonen:

$$F_{H_S}\left(h_S\right)=\Phi\left(z_1\right)$$
$$F_{T_P|H_S}\left(t_P \mid h_S\right)=\Phi\left(z_2\right) \quad \text{(J)}$$

Her er ϕ den kumulative fordelingen til en standard normal variabel, z_1 og z_2 er og standard normalfordelt. Ligningene i (J) gir videre:

$$h_S=-\ln\left(1-\Phi\left(z_1\right)\right)$$
$$t_P=\frac{-1}{h_S}\cdot\ln\left(1-\Phi\left(z_2\right)\right) \quad \text{(K)}$$

Vi vet og at i normal-rommet kan konturlinjen finnes som en sirkel der radien, r, er avhengig av ønsket sannsynlighetsnivå. Langs konturlinjen har vi at $z_1^2 + z_2^2 = r^2$. I dette eksempelet er "r" satt til 1,5, og alle punktene i normal-rommet er så transformert tilbake til H_S-T_P rommet ved hjelp av ligningene i (K). Denne konturlinjen man får da er vist som den blå linjen i figur A.

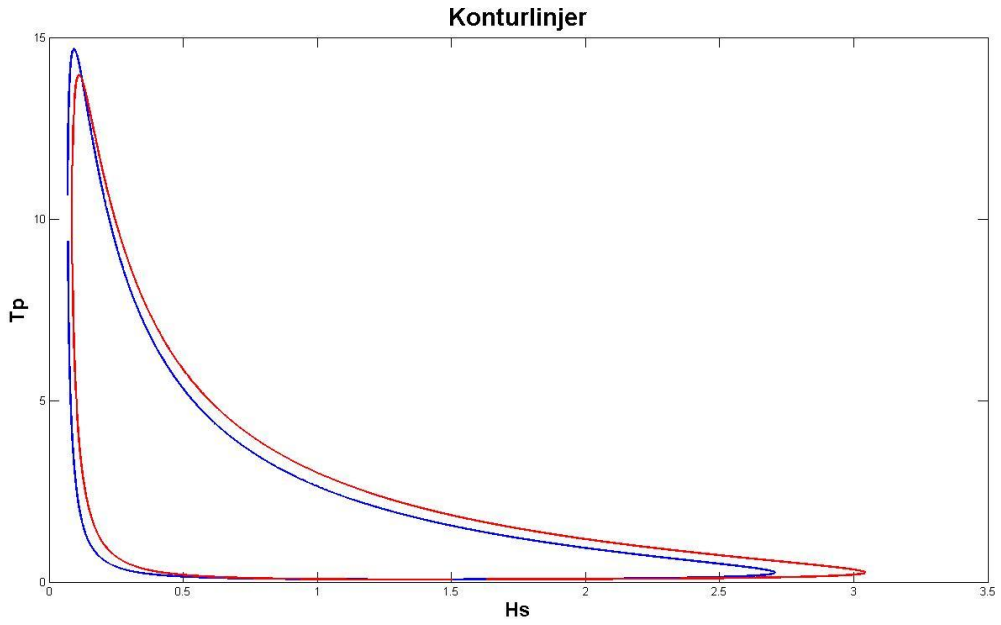
Konturlinje nr. 2 er gitt ved transformasjonen:

$$F_{T_P}\left(t_P\right)=\Phi\left(z_1\right)$$
$$F_{H_S|T_P}\left(h_S \mid t_P\right)=\Phi\left(z_2\right) \quad \text{(L)}$$

Her er ϕ den kumulative fordelingen til en standard normal variabel, z_1 og z_2 er og standard normalfordelt. Ligningene i (L) gir videre:

$$t_P=\frac{1}{1-\Phi\left(z_1\right)}-1$$
$$1-e^{-h_S\left(1+t_P\right)}\left(1+\left(1+t_P\right)h_S\right)=\Phi\left(z_2\right) \quad \text{(M)}$$

Det er vanskelig å finne et analytisk uttrykk for h_S ved hjelp av nederste ligning i (M), denne løses derfor numerisk i Matlab for gitt t_P og z_2. For hvert punkt av z_1 og z_2 langs sirkelen i normal-rommet med radius 1,5 finnes korresponderende punkt i H_S-T_P rommet. Denne konturlinjen man får da er vist som den røde linjen i figur A.



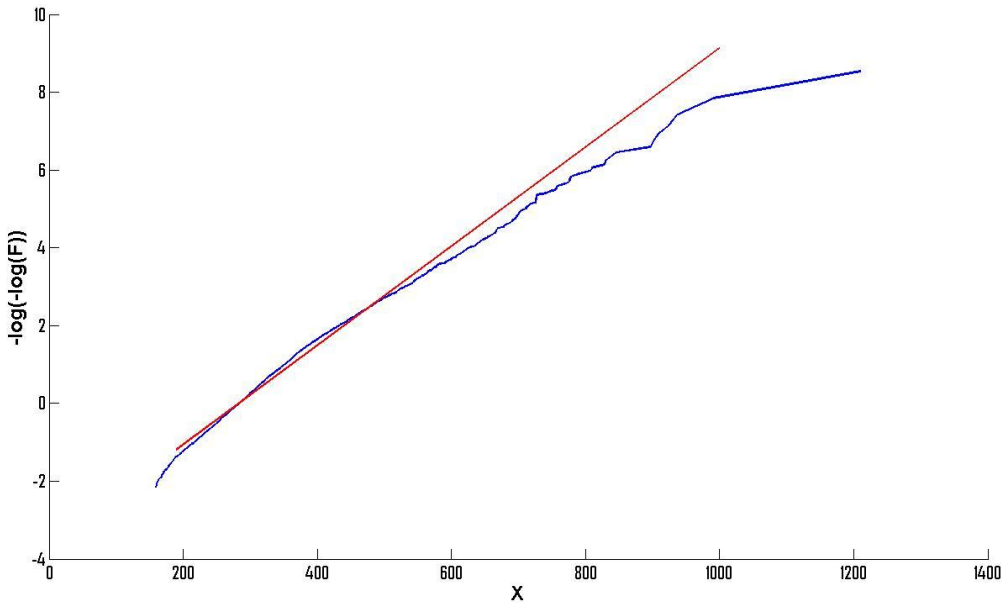
Figur A: Konturlinjer. Blå linje tilhører transformasjon nr. 1, rød linje tilhører transformasjon nr. 2.

Som man ser av figur A er ikke linjene så veldig forskjellig, men det er allikevel en viss forskjell som i noen tilfeller kanskje kan være viktig. Generelt kan man si at ulikheten mellom de to linjene blir mindre jo nærmere de to fordelingene (som skal transformeres) er normalfordelt. Forskjellen vil også minke jo større radien i normal-rommet er, (Haver og Winterstein (2010)).

C. Metode 2 og 3 ved Ψ = 2

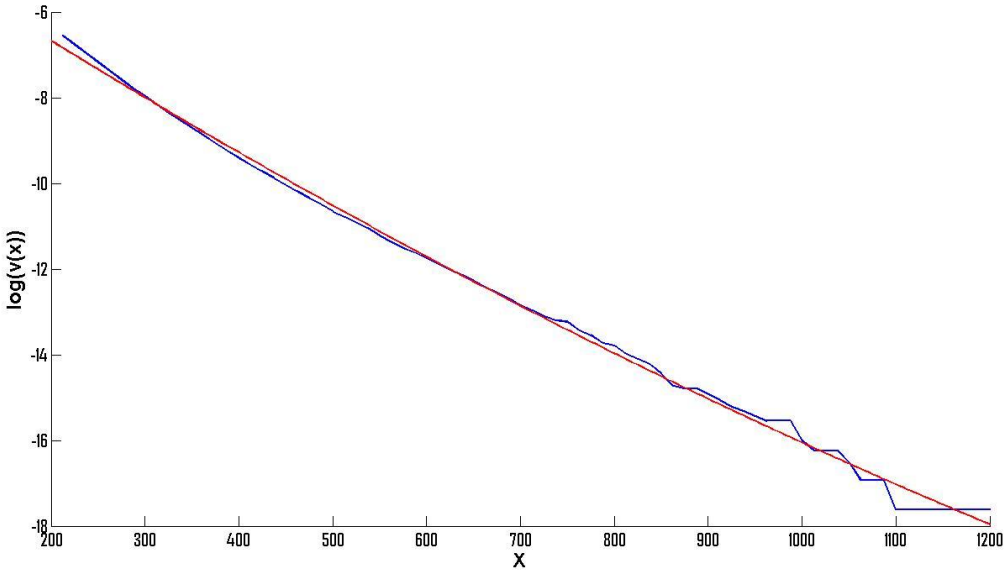
I dette avsnittet er de resulterende grafene med tilpassede modeller for metode 2 og 3 vist når Ψ = 2. Det er ikke tatt med resultatene for Ψ = 1,5 fordi disse er en mellomting av hva man finner når Ψ = 1 og Ψ = 2, og er derfor ikke så interessant.

I figur B er årlige maksima plottet i et Gumbelpapir når Ψ = 2 (blå linje). Rød linje viser en modell tilpasset ved momentmetoden. Som man ser av figur B er den blå linjen ganske krum, og Gumbelantagelsen er derfor ikke god når Ψ = 2. Dette kan nok forklare de avvikende resultatene i tabell 5 med denne metoden.



Figur B: Årlige maksimum i Gumbelpapir, Ψ = 2 (blå linje). Rød linje er en tilpasset modell ved momentmetoden (se kapittel 3.1.).

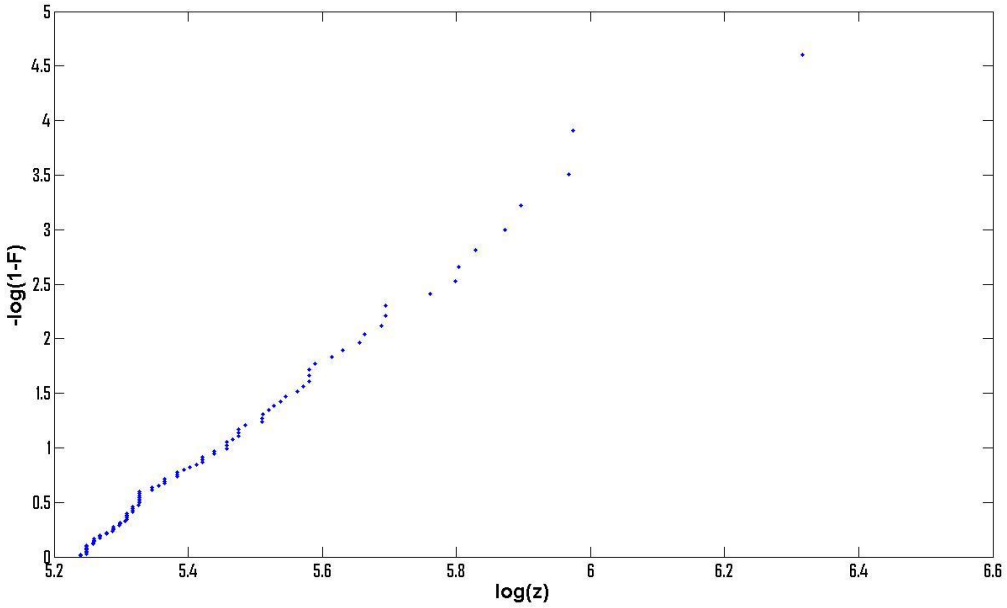
I figur C er logaritmen til nivåoppkrysningsfrekvensen, $\nu^+(x)$, plottet når Ψ = 2 (blå linje). Man kan se at $\log(\nu^+)$ plottet mot x krummer. Ved kurvetilpasning i Matlab tilpasses en funksjon for $\nu^+(x)$ på formen gitt i ligning (16). Tilpasset funksjon er vist som rød linje i figur C.



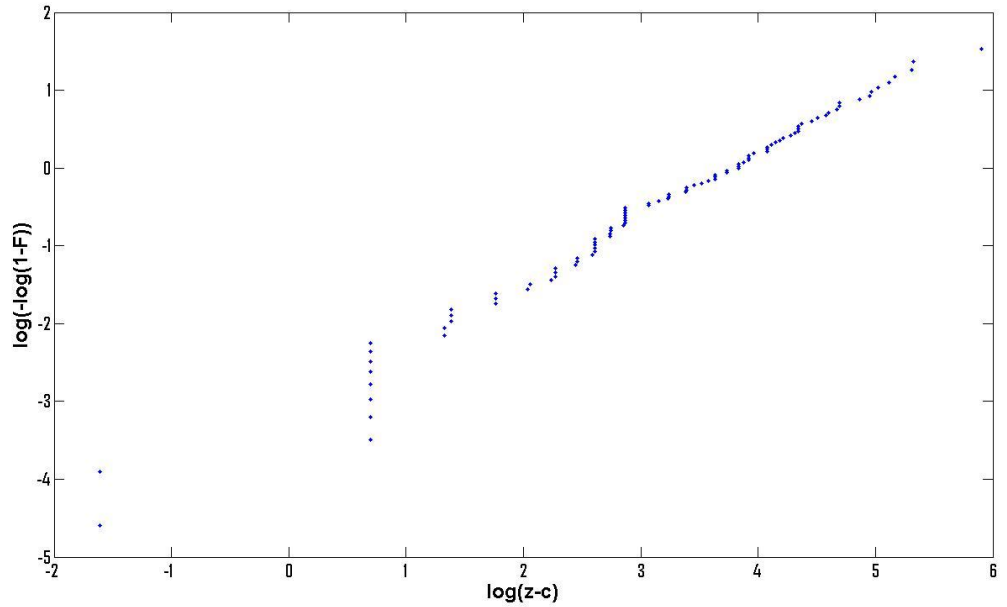
Figur C: log($v^+(x)$) som funksjon av nivået x når Ψ = 2.

D. Følsomhet ved metode 4

I figur D og E nedenfor er observasjonene av Z plottet i et Pareto og et Weibullpapir.

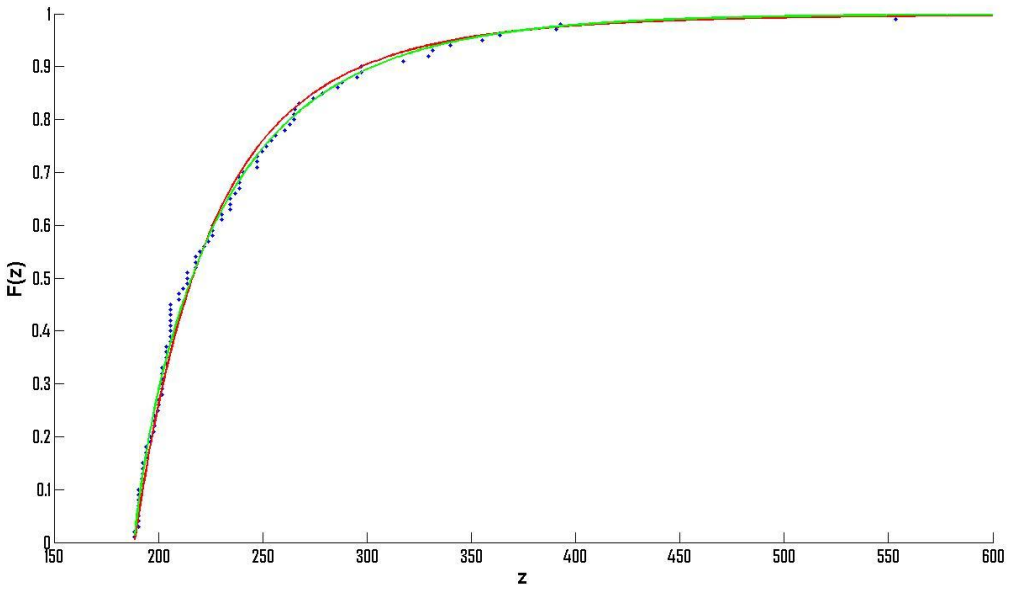


Figur D: Observasjonene av Z plottet i et Paretopapir ($H_{S,storm}$ = 10m og Ψ = 2).



Figur E: Observasjonene av Z plottet i et Weibullpapir ($H_{S,storm}$ = 10m og Ψ = 2).

Som man ser av de to figurene over kan både en Pareto og en Weibullfordeling være gode alternativer, skulle man tro i alle fall. En Paretofordeling og en Weibullfordeling er tilpasset, og plottet i figur F nedenfor.



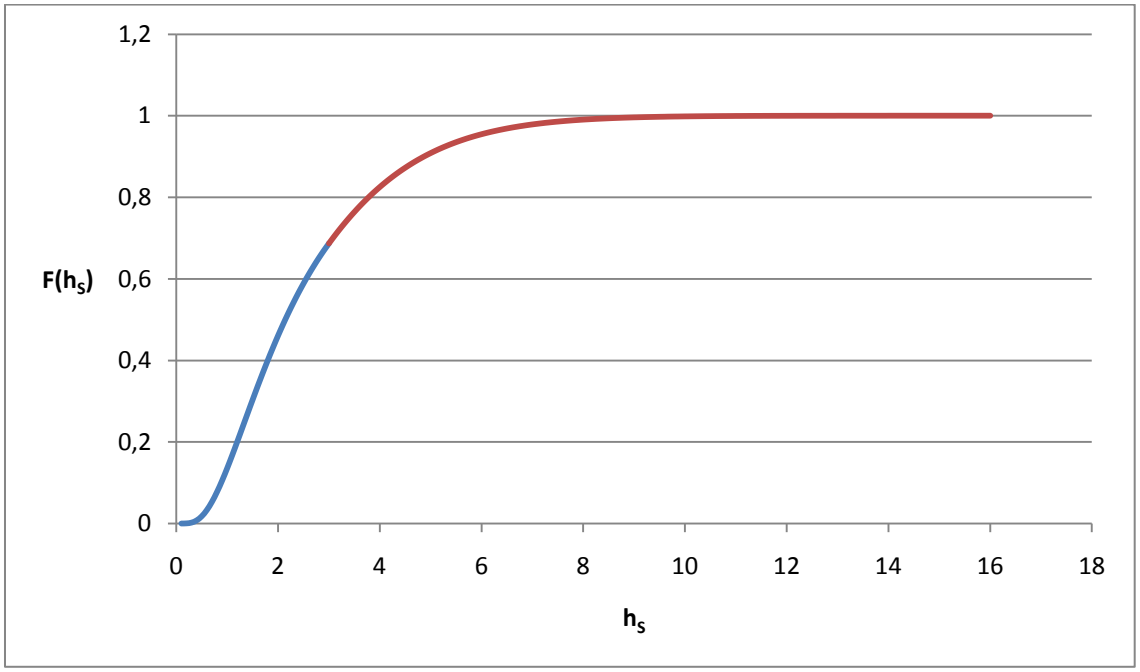
Figur F: Tilpasset Paretofordeling (rød linje), tilpasset Weibullfordeling (grønn linje) og observasjoner av Z (blå prikker).

Pareto og Weibullfordelingen ser like ut, men forskjellene i resultatene blir veldig stor avhengig av hvilken av de to fordelingene man velger å benytte, se kapittel 5.3.

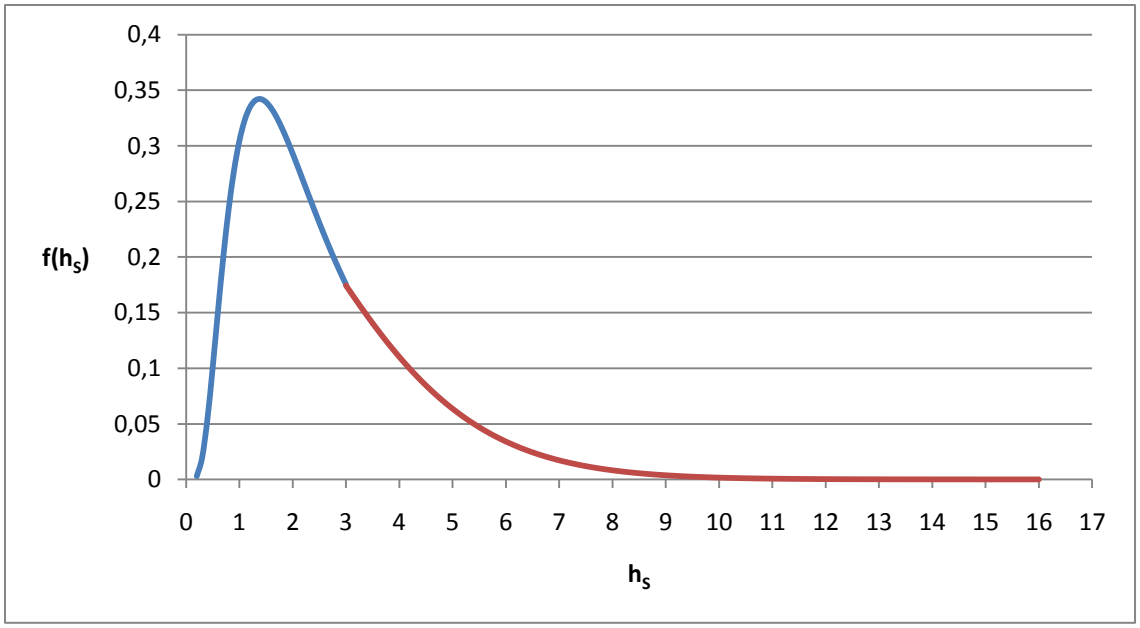
E. Fordeling til H_S

Det er benyttet to fordelinger for H_S, en for $H_S < 3$ meter og en for $H_S \geq 3$ meter, se kapittel 2.2.

I figur G og H nedenfor er den kumulative fordelingen og tetthetsfunksjonen til H_S plottet med fordelingen gitt i ligning (6) og (7). Fargen er blå for $H_S < 3$m og rød for $H_S \geq 3$m.



Figur G: Kumulativ fordeling til H_S som gitt i ligning (6) og (7). Den blå linjen viser nedre del av fordelingen til H_S (lognormaldelen), mens den røde linjen viser øvre del av fordelingen (Weibulldelen).



Figur H: Tetthetsfunksjonen til H_S gitt i ligning (6) og (7). Den blå linjen viser nedre del av fordelingen til H_S (lognormaldelen), mens den røde linjen viser øvre del av fordelingen (Weibulldelen).

F. Kopulaer

En kopula er et alternativ til en ”vanlig” multivariabel fordelingsfunksjon. En kopula er egentlig bare en fordeling til uniforme (0,1) variabler. Hvis man for eksempel vil modellere variablene X og Y med en kopula må man transformere disse til uniforme variabler, U og V, ved hjelp av de respektive marginalfordelingene (sannsynlighets integral transformasjon).

$$\begin{aligned} F_X\left(X\right) &= U \\ F_Y\left(Y\right) &= V \end{aligned} \tag{N}$$

Man kan da finne den bivariate fordelingen til U og V, dette er en kopula. For variablene X og Y med fordeling $F_{X,Y}(x,y)$, og med marginalfordelinger $F_X(x) = F_{X,Y}(x,\infty)$ og $F_Y(y) = F_{X,Y}(\infty,y)$ vil det finnes en kopula (Sklar’s teorem), C, slik at:

$$P\left(X < x, Y < y\right) = F_{X,Y}\left(x,y\right) = C\left(F_X\left(x\right), F_Y\left(y\right)\right) = C\left(u,v\right) \tag{O}$$

Her er C kopulafunksjonen, $F_X(x) = u$ og $F_Y(y) = v$. For mer om kopulaer se for eksempel (Nelsen (2006)).